Corrections to
Bias Calibration in Finite Population Estimation

1. Page 1: "Chambers (1994)" should be replaced by "Welsh and Ronchetti (1994)".

2. Page 2: "Dorman (1990)" should be replaced by "Dorfman (1993)".

3. Page 3: "(details refer to the paper)" is replaced by "(equation (1.3) of page 3)".

4. Page 7: Sec. 2.2 To the end of the second paragraph, add "This is confirmed by a high positive correlation between ROOMS and BEDROOMS as well as between ROOMS and BATHROOMS."

5. Page 8: At line 5, after "...the 1991 Brazilian Population Census for all the households." insert "Whereas the Total Monthly Income was unknown for ninety percent of the households."

6. Page 24: In addition to the definitions of Σ, Σ_{11} and Σ_{22}, add

$$\Sigma_{12} = \begin{pmatrix} \sigma_{1,n+1} & \cdots & \sigma_{1,N} \\ \vdots & \ddots & \vdots \\ \sigma_{n,n+1} & \cdots & \sigma_{n,N} \end{pmatrix} \quad \text{and} \quad \Sigma_{21} = \Sigma_{12}'$$

7. Page 26: Next to $\eta(x) = (1-x^2)^2 I(|x| \le 1)$ add $c_E = 4.685$.

8. Page 37: At the end of the second paragraph of Section 4.2, add "Of course there is also a category of both positive and negative outliers in the sample under the model-based framework. However, in such a case, if the number and magnitude (i.e. their combined effect) of positive outliers is not significantly different from the number and magnitude or their combined effect of negative outliers, influence from each on the estimates tend to cancel out each other and this situation is similar to the case of Sample 1. This argument was confirmed by the result of a study of such a sample actually carried out in the work of this thesis but for brevity reason, its result was excluded in this thesis.

9. Page 38: At the end of the first paragraph of Section 4.3, add "Certainly this choice of optimal c depends on knowing the population total of TMI and hence results for estimators based on such "optimal" values are more of theoretical interest. Note that figures with bold typeface indicate either they are the true total or they are the closest estimates to the true total within their class of estimates."

10. Page 38: Follow the first paragraph of Section 4.3, add this paragraph

> "Results of the following tables are the estimates of population total given by the various estimators described in Section 4.1, the figures shown here are aggregated and then exponentiated so that they are presented on the raw scale for the convenience of the reader."

11. Page 40: At the end of the second paragraph, insert the following paragraph

> "It is worthwhile to point out that the strong performance of the Number-raised estimaor when compared with all the regression estimators based on model II in Tables 4.1 and 4.2 may suggest that no amount of robust fitting can help a very poor model."

12. Page 41: At line 3 of the first paragraph, the phrase "which can even be made to hit the true total." is replaced by "which can be made to equal to the true total ."

13. Page 42: At the beginning of Section 4.5 add this paragraph

> "Based on the results of this section, a strategy of estimating the population total is suggested. However, due to the limited nature of this numerical study, its usefulness is subjected to further investigation."

14. Page 51: At line 2 of Section 5.1, "..described in chapter can be.." should be "..described in chapter 3 can be..".

15. Page 53: At the end of the first paragraph of Section 5.2, "The performance of each estimator is presented." is replaced by "The performance of each estimator is presented by examining their corresponding quantile estimates. As indicated by Chambers and Dunstan (1986), in practice, there are times that the primary aim of the survey is to identify subgroups in the population whose values for particular variables lie substantially below or above a certain value."

16. Page 70: In between the references of Rao, C.R. (1971) and Searls, D. T. (1966) add the reference "Royall, R.M. (1970). On finite population sampling under certain linear regression models. *Biometrika* **57** 377-387."

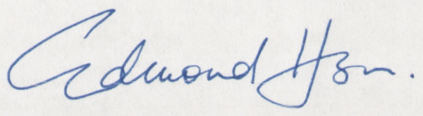
Lung Keung Hsu
April 1996

BIAS CALIBRATION IN FINITE POPULATION ESTIMATION

A thesis submitted for the degree of
Master of Statistics
of the Australian National University
July 1995

Lung Keung Hsu

I, Lung Keung Hsu, declare that all work in this thesis is my own, supervised by Dr. Alan Welsh. All references to the work of others have been duly acknowledged.



THE AUSTRALIAN
LIBRARY
NATIONAL UNIVERSITY

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Alan Welsh, for his helpful comments throughout both the research and writing stages of this thesis. He has been a consistent source of encouragement, criticism and guidance, without whose patience and support this thesis would not have been possible. I am also grateful to Dr. R. L. Chambers for making the Brazilian data available to me. Thanks also to Professor Chip Heathcote for his encouragement and general support.

ABSTRACT

In this thesis, the role of bias calibration in finite population estimation is investigated by using a set of household survey data. Expressions of some bias calibrated estimators for estimating the population total and population distribution function under a general linear model framework are derived. A superpopulation model is first found for the household survey data then the various bias calibrated estimators are applied. Results of the bias calibrated estimation are presented and analysed.

CONTENTS

1. Introduction

Survey samplers are often faced with the problem of estimating finite population parameters from a sample containing extreme values or representative outliers. Chambers (1986) clearly described the difference between representative outliers and nonrepresentative outliers. Representative outliers are sample elements which are extreme values relative to the bulk of the data but are correctly recorded and therefore we have no good reason to assume there are no more similar outliers in the nonsample portion of the population. Nonrepresentative outliers are sample elements whose values are in some sense unique, for example, due to incorrect observation and he suggested that nonrepresentative outliers problem could be dealt with by survey editing and imputation methods.

As pointed out by Chambers (1994), when the number of representative outliers is substantial, post-stratification methodologies can be employed. See for example Glasser (1962), Kish (1965), Searls(1966), Rao (1971) and Hidiroglou and Srinath (1981). But when there are only a relatively few representative outliers, the appropriate weights to be assigned to the units are difficult to determine.

At present, there are basically three distinct approaches to dealing with the problem of a sample containing a few representative outliers. Chambers and Kokic (1993) made explicit the "modified weights approach" which reduces the sample weights associated with sample outliers, leaving their Y-values unchanged where Y is the value of interest. Another approach is the "modify values approach" which modifies the Y-values associated with the sample outliers, leaving their weights unchanged. The remaining approach is outlier robust estimation and this approach is the emphasis of this thesis.

Chambers (1986) introduced the use of outlier robust estimator of the population total under a simple linear regression superpopulation model. This outlier robust estimator is basically the sum of the sample total, an M-estimate of the nonsample total and a robust estimate of the bias. The idea of this estimator was developed and extended in Chambers and Kokic (1993) where it is referred to as the bias calibrated estimation of the population total. To quote Chambers and Kokic (1993), "That is, one first estimates this finite population total as if the working model applies to all nonsample units. This 'working model' estimator is generally

biased if sample outliers are representative. A bias calibration term is then applied which uses the information in the sample outliers to compensate for the bias in this initial working model estimator".

Welsh and Ronchetti (1994) clarified the motivation for this bias calibrated estimator of the population total and showed that bias calibration is essential in constructing estimators of finite population parameters. This paper linked the problem of total estimation to distribution function estimation and proposed a methodology based on the use of robust estimates and a bias calibrated form of the Chambers and Dunstan (1986) estimator of the distribution function. This idea led to an alternative bias calibrated estimator of the population total to that of Chambers (1986). One more significant contribution of this paper is that it has discovered a very simple yet efficient strategy to estimate the quantile function of the beef farm data. This data was considered by Dorman (1990), Chambers, Dorfman and Wehrley (1993) and Chambers and Kokic (1993). The population distrbution function estimator proposed by Welsh and Ronchetti (1994) is (details refer to the paper)

$$\hat{F}(t,c) = \frac{n}{N} \cdot F_1(t) + \frac{N-n}{N} \cdot \hat{F}_2(t,c), \qquad t \in R$$

where

$$\hat{F}_2(t,c) = \frac{1}{n(N-n)} \cdot \sum_{j=n+1}^{N} \sum_{i=1}^{n} \mathrm{I}\left\{ \hat{\beta}_R X_j + v^{\frac{1}{2}}(X_j) \cdot c\hat{\sigma}\psi\left\{ \frac{Y_i - \hat{\beta}_R X_i}{c\hat{\sigma} v^{\frac{1}{2}}(X_i)} \right\} \le t \right\}$$

The strategy is to change the bias calibration c over the support of the distribution in such a way that c increases as we move into the right tail. This strategy works extremely well for the beef farm data and seem to be a very good population distribution function estimator with great potential for general use.

The work of this thesis is an exploration of the role of bias calibration in the robust estimation of the population total and population distribution function. The data used for this analysis is the Brazilian data which is a subset of the data collected during the Test Population Census of Limeira, 1988. This data set contains household characteristics for 954 households. Interest is in the relationship between household total income and other household

characteristics. Firstly, a superpopulation model is found for the Brazilian data. Secondly, various population total and population distribution function estimators described in Welsh and Ronchetti (1994) are extended so that they can be applied under a general linear model framework. Then the performance of these estimators applied to the Brazilian data are compared and discussed so as to gain some insight into the effect of different choices of the bias calibration on the estimation of population total and distribution function in general.

Note that in order to concentrate on the investigation of the role of bias calibration in estimating finite population parameters, throughout the work of this thesis we assume that the process used to decide which population elements to include in the sample was independent of the population values of interest (ignorable sampling) and that any sample outliers in the sample selected are representative outliers.

2. The Superpopulation Model for the Brazilian Data

2.1 *The Brazilian Data*

The Brazilian data used for the analysis is a subset of the data collected during the Test Population Census of Limeira, 1988. This was a pilot census carried out by IBGE (the Brazilian Institute for Geography and Statistics) to rehearse the methodology and procedures planned for the 1991 Population Census.

The 1988 Test Census was carried out in 2 waves. In the first visit, all the households were visited by an interviewer who used a short questionnaire, called "Basic Questionnaire", to collect data on a handful of characteristics of the household which are sex, age, education and "proxy total income" of the heads of households. For other individuals, only sex, age, relationship to head of household and literacy were collected. In a second wave of data collection, a more detailed questionnaire, called "Sample Questionnaire", was collected from a 10% sample of households selected systematically within each enumeration area. The Sample Questionnaire contained all the questions in the Basic Questionnaire plus many others, with more detailed information about households and individuals. As an example, income by source (6 possible sources) was collected from every individual aged 10 or over. The "Total Monthly Income" was derived by adding the incomes from these various sources. In the 1991 Population Census, the practice of two visits to the households was not adopted, but still the "Basic Questionnaire" was collected from around 90% of the households, and the "Sample Questionnaire" from the remaining 10%, meaning that information for the variables collected in the "Basic Questionnaire" are available for 100% of households.

The Brazilian data used here as our population is a subset of the data collected in the second wave during the 1988 Test Census. This data set contains one record for each household enumerated using the Sample Questionnaire in enumeration areas 1 to 40 in Limeira, totalling 954 records. The records were labelled sequentially from 1 to 954, and all identification information was removed to avoid breach of confidentiality. These households were selected by random systematic sample within each enumeration area.

The information contained in the data set includes 2 Identification Variables, 11 Households Variables and 15 Head of Households Variables. The variables are

	NAME	DESCRIPTION
Identification Variables	1. LABEL	label of household
		range : 1 - 954
	2. EA	enumeration area
		range : 1 - 40
Household Variables	1. HOUSE	indicator that building type is house
		0 = flat or other type of building
		1 = house
	2. OWNED	indicator that building is owned by occupants
		0 = rented or other
		1 = owned
	3. ROOMS	number of rooms in household
		range : 1 - 18
	4. BEDROOMS	number of bedrooms in household
		range : 1 - 5
	5. BATHROOMS	number of bathrooms in household
		range : 0 - 5
	6. FILTER	indicator of water filter in household
		0 = no water filter in household
		1 = water filter in household
	7. TV_BW	indicator for B&W TV in household
		0 = no B&W TV in household
		1 = B&W TV in household
	8. TV_COLOR	indicator for color TV in household
		0 = no color TV in household
		1 = color TV in household

	9. CAR	indicator of car in household 0 = no private car in household 1 = at least one car in household
	10. TELEPHONE	indicator of telephone in household 0 = no telephone 1 = at least one telephone
	11. WASHING	indicator of washing machine in household 0 = no washing machine in household 1 = washing machine in household
Head of Household Variables	1. SEX	indicator that head of household is male 0 = head of household is female 1 = head of household is male
	2. AGE	age in years range : 17 - 93
	3. LITERACY	indicator that head of household can read & write 0 = cannot read or write 1 = can read and write
	4. YEAR_EDU	number of years in education range : 0 - 17
	5. P_INCOME	proxy of total income range : 0 - 2,400,000
	6. WHITE	indicator that head of household is of white race 0 = not white 1 = white
	7. MARRIED	indicator that head of household is married 0 = not married 1 = married
	8. HOURWORK	number of hours worked weekly on main occupation by head of household range : 0 - 98 (0 = no work)

9. INCOME1 fixed monthly income from main occupation
range : 0 - 1,600,000

10. INCOME2 variable monthly income from main occupation
range : 0 - 2,400,000

11. INCOME3 monthly income from other occupation
range : 0 - 600,000

12. INCOME4 monthly income from retirement
range : 0 - 800,000

13. INCOME5 monthly income from pension
range : 0 - 700,000

14. INCOME6 other monthly income
range : 0 - 3,000,000

15. INCOME total monthly income
range : 0 - 3,131,000
(sum of INCOME1 to INCOME6)

2.2 *Exploratory Data Analysis of the Brazilian Data*

For the Brazilian data, interest is in trying to understand how the "Total Monthly Income" (TMI) depends on the other variables. As the Total monthly income is the sum of the six income sources, the six different income sources are not taken as explanatory variables.

In order to reduce the number of explanatory variables, the two variables BEDROOMS and BATHROOMS are excluded but the variable ROOMS is retained since ROOMS is the total number of rooms in household and should convey the same information as BEDROOMS and BATHROOMS.

As a result, eighteen variables are left as predictors. As shown in Figure 2.1, the distributions of Total Monthly Income and the Proxy Income (PI) are highly skewed as expected, thus the log transformation is taken for these two variables. It is noteworthy that the log Proxy Income variable is very highly correlated with the log Total Monthly Income, as high

as 0.9019. The table 2.1 shows the four highest correlations greater than 0.5 in absolute value among the eighteen explanatory variables and the response variable.

Table 2.1

Variables	Correlation
log TMI & log PI	0.9019
SEX & MARRIED	0.7803
AGE & HOURWORK	-0.6177
BWTV & COLORTV	-0.5958

Hence the Proxy Income serves as a very useful benchmark in predicting the Total Monthly Income. Usually such a benchmark is rare in survey sampling, but it was available in the 1991 Brazilian Population Census for all the households. Figure 2.2 shows a scatter plot of log TMI versus log PI. There are a number of points which are very far away from the rest of the population and deserve special attention and they can be grouped into three categories as :

Category I : Zero Total Monthly Income and Zero Proxy Income											
Household Identification	39	86	276	322	544	545	672	763	776	805	909

Category II : Zero TMI and Non-zero PI		
Household Identification	166	941

Category III : Non-zero TMI and zero PI	
Household Identification	443

Clearly the households of Category I successfully foretold their Total Monthly Income to be zero (even though we do not know how they could survive) while Categories II and III tell us that households 166, 941 and 443 completely miss-predicted their Total Monthly Income. Figure 2.3 shows a scatter plot of log TMI versus log PI without those households mentioned in the three categories which displays a strong positive linear relationship between log TMI and log PI.

Relationships between log TMI and three other predictors, AGE, HOURWORK and YEAR_EDU (interval nature) are also shown by the scatter plots in Figure 2.4. No particular pattern can be recognised from these three scatter plots.

2.3 *Linear Regression Model obtained by Robust Regression*

Our objective here is to find a simple linear model which describes the behaviour of log Total Monthly Income. As has been mentioned before, eighteen explanatory variables are used to form the "Full Model" which are EA, HOUSE, OWNED, ROOMS, FILTER, TV_BW, TV_COLOR, CAR, TELEPHONE, WASHING, SEX, AGE, LITERACY, YEAR_EDU, log P_INCOME, WHITE, MARRIED, and HOURWORK respectively.

In general, if Y is the response variable and $X_1,\ldots,X_p$ are the predictors, a linear regression model is formally defined as

$$\mathrm{Y} = \mathrm{X}\beta + e \qquad (2.1)$$
$$\mathrm{var}(e) = \Sigma$$

such that

$$\mathrm{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad e = \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}$$

$$\mathrm{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1N} \\ \vdots & \ddots & \vdots \\ \sigma_{N1} & \cdots & \sigma_{NN} \end{pmatrix}$$

where Σ is a symmetric positive definite matrix whose value is not necessarily known.

For our case, Y is the log TMI and $X_1, X_2, \ldots, X_{19}$ are the predictors (including the intercept). The estimate of β here, say $\hat{\beta}$ is the M-estimate (Huber 1981) obtained via robust regression using the Huber Ψ - function so that $\hat{\beta}$ is the solution to the equation

$$\sum \Psi\left(\frac{y_i - x_i\hat{\beta}}{d}\right) = 0 \quad \text{such that} \qquad \Psi(x) = \begin{cases} -c & x < -c \\ x & -c \le x \le c \\ c & c < x \end{cases} \qquad (2.2)$$

where $d = \dfrac{median\left|\left(y_i - x_i\hat{\beta}\right) - median\left(y_i - x_i\hat{\beta}\right)\right|}{0.6745}$ and $c = 1.345$

Figure 2.5 shows the diagnostics for this full model. The households 166, 443 and 941 are the most extreme outliers. If these three households are excluded, the model seems to be a reasonable one.

To achieve parsimony, backward elimination was employed to find the smallest possible subset of explanatory variables which best describes the log Total Monthly Income. The procedure was to drop the least significant predictor from the model each time according to the partial F-test statistics. CAR, was found to be the first one to go, then COLOR_TV, WHITE, ROOMS, LITERACY, WASHING, AGE, TELEPHONE, OWN, BW_TV, YEAR_EDU, SEX, MARRIED, FILTER, HOUSE and EA sequentially. At the end, **log PI** and **HOURWORK** were found to be the most important predictors and were taken as the predictors for the final model. Table 2.2 gives the results of the robust estimates of the parameters, their standard errors and their t-statistics.

Table 2.2

	Value	Standard Error	t-statistic
Intercept	0.01258	0.006985	1.8
log Proxy Income	0.99899	0.000653	1528.9
Hours of Work	0.00018	0.000045	4.0

The t-statistic 1528.9 of log Proxy Income shows that log Proxy Income accounts for most of the explanation of the behaviour of log Total Monthly Income. Diagnostics of this final model shown in Figure 2.6 are very much the same as those in Figure 2.5. In order to examine the adequacy of this model without the influence of the three extreme outliers (households 166, 443, and 941) and also the eleven zero TMI - zero PI households, this final model is fitted again to the remaining 940 observations and the diagnostics of this fit are shown in Figure 2.7. Figure 2.8 is an enlarged version of the residual plot of this fit. To our surprise,

heterocedasticity does not seem to appear. This may be a result of the log transformations applied to both the Total Monthly Income and the Proxy Income.

In any normal sample survey situation, the availability of a benchmark value with performance such as the Proxy Income is rare. So the Brazilian data is a very unusual case. In order to broaden the applicability of our results from our analysis to the Brazilian data to more real life situation, we also fitted a linear model by robust regression without the log Proxy Income.

Procedures to search for this model were exactly the same as before. The final model we obtained included predictors OWN, ROOMS, COLOR_TV, CAR, TELEPHONE, SEX, YEAR.EDU and HOURWORK. Table 2.3 presents the results of the parameter estimates.

Table 2.3

	Value	Standard Error	t-statistic
Intercept	8.2428	0.14863	55.46
Own	-0.1216	0.04877	-2.49
Rooms	0.0892	0.01074	8.30
Color_TV	0.1348	0.05975	2.26
Car	0.3567	0.05238	6.81
Telephone	0.3147	0.05347	5.89
Sex	0.3158	0.05939	5.32
Year.Edu	0.0697	0.00560	12.45
Hourwork	0.0149	0.00105	14.25

Diagnostics for this model (Figure 2.9) clearly indicate that the fit is reasonable except for the thirteen households which have zero Total Monthly Income. That is to say the model without Proxy Income is completely unable to account for the situation where a household has no income at all. Figure 2.10 shows that heterocedasticity does not seem to be a problem for this model either.

As a result, for the Brazilian data, we would consider two superpopulation models. One is the model with log Proxy Income and Hourwork, the other is the model without log Proxy Income but eight other predictors. Formally the models are

Model I (with Proxy Income)

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + e \qquad (2.3)$$

$E(e) = 0$ and $\text{var}(e) = \sigma^2 I$ $I: N \times N$

Y, X_1, X_2 and e are both $N \times 1$ vector with $N = 954$

$\beta_0, \beta_1, \beta_2$ and σ^2 are parameters to be estimated by robust regression

Y = log Total Monthly Income

X_1 = log Proxy Income X_2 = Hours of work

Model II (without Proxy Income)

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_8\beta_8 + e \qquad (2.4)$$

X_1 = Own X_5 = Telephone

X_2 = Rooms X_6 = Sex

X_3 = Color_TV X_7 = Years of Education

X_4 = Car X_8 = Hours of Work

and the rest are the same as in Model I.

Figure 2.1 : Boxplots before and after log transformation

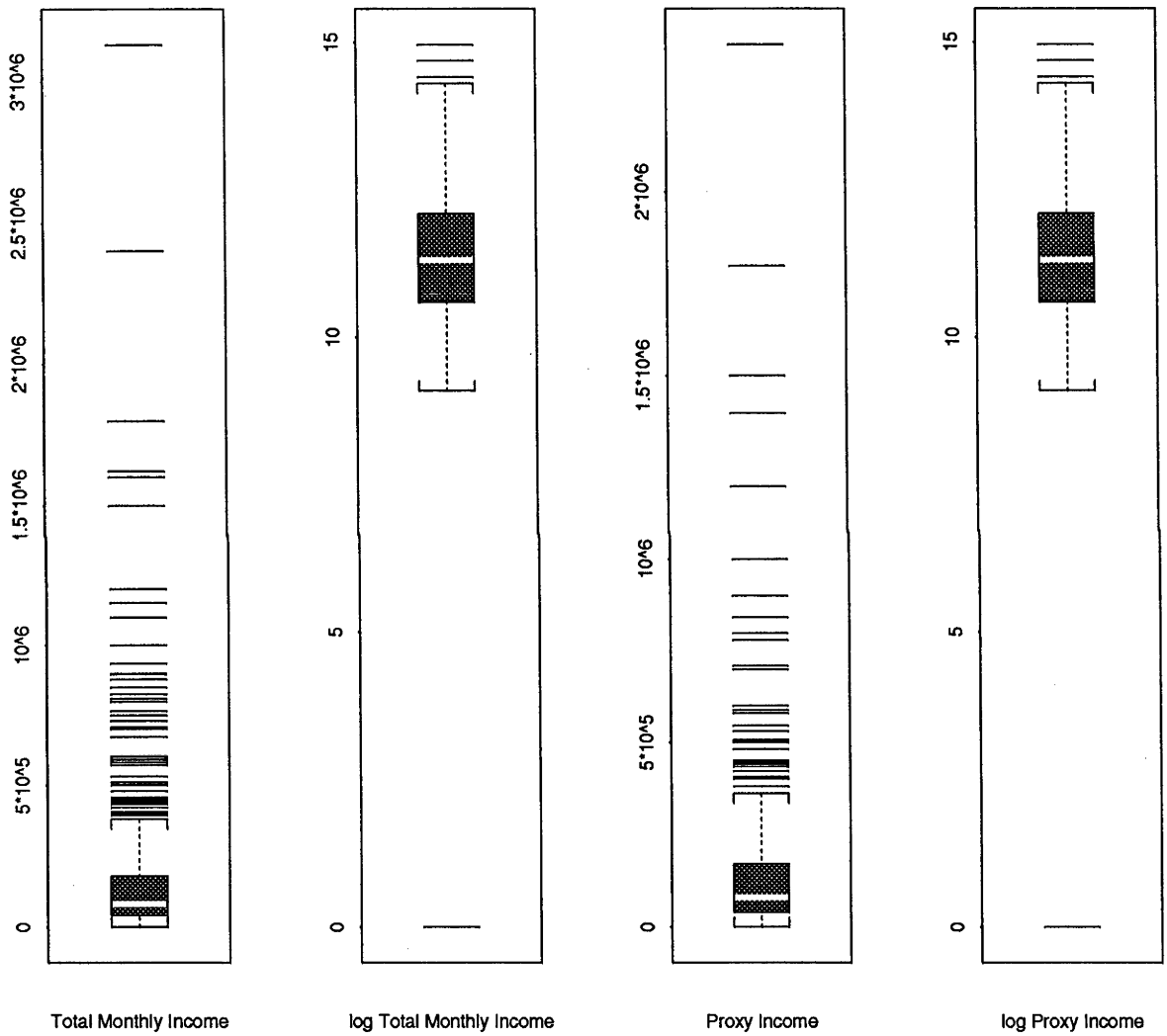


Figure 2.2 : Scatter Plot of log TMI vs log PI for Population

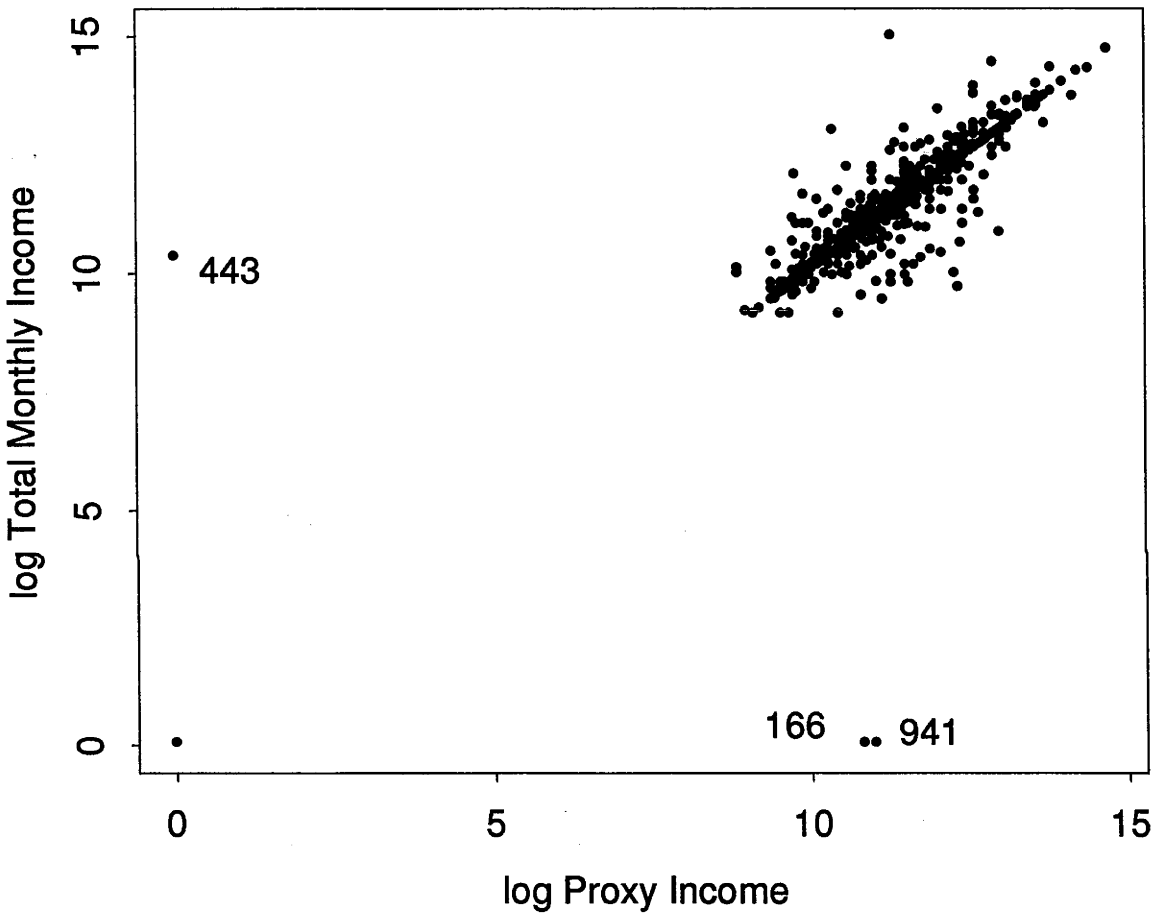


Figure 2.3 : Scatter Plot of log TMI vs log PI without zero PI nor zero TMI

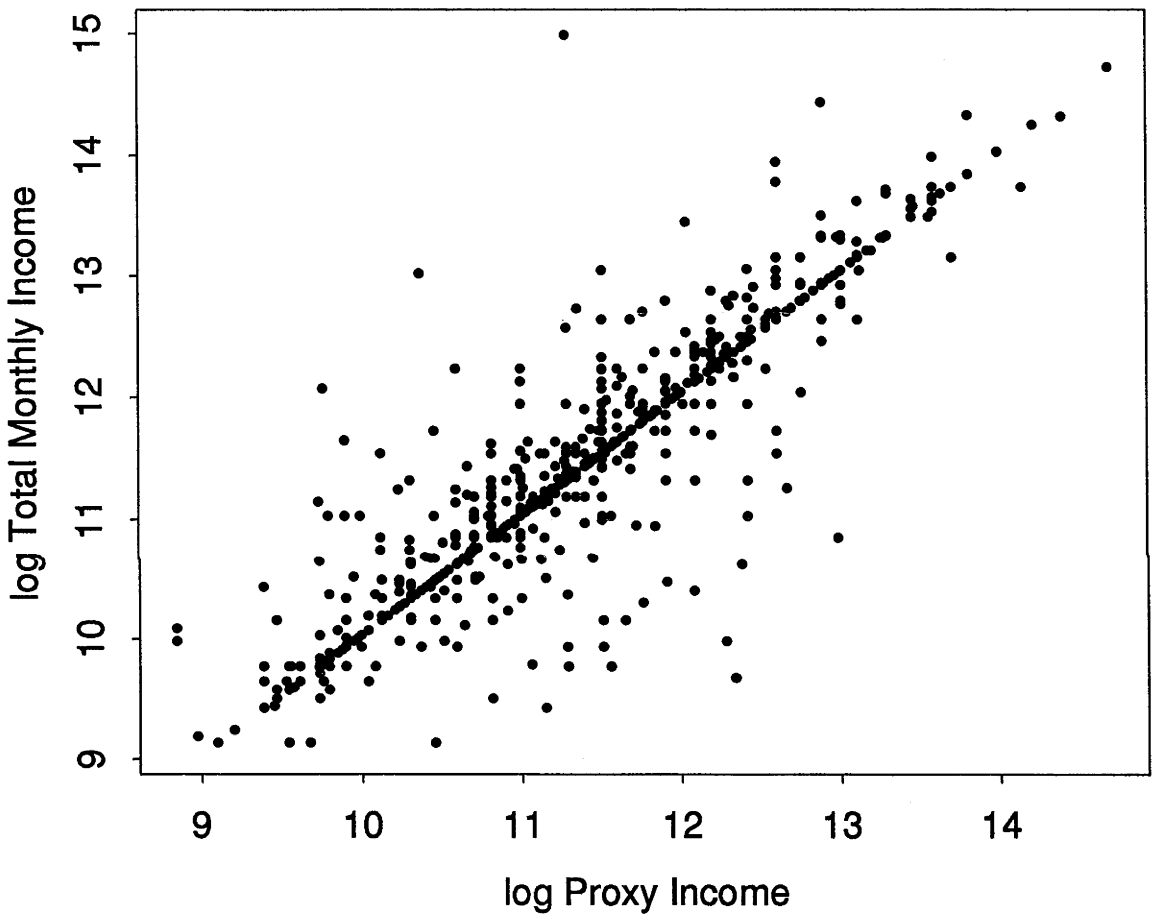


Figure 2.4 : Scatter Plots of log TMI vs AGE, YEAR of EDUCATION & HOURWORK

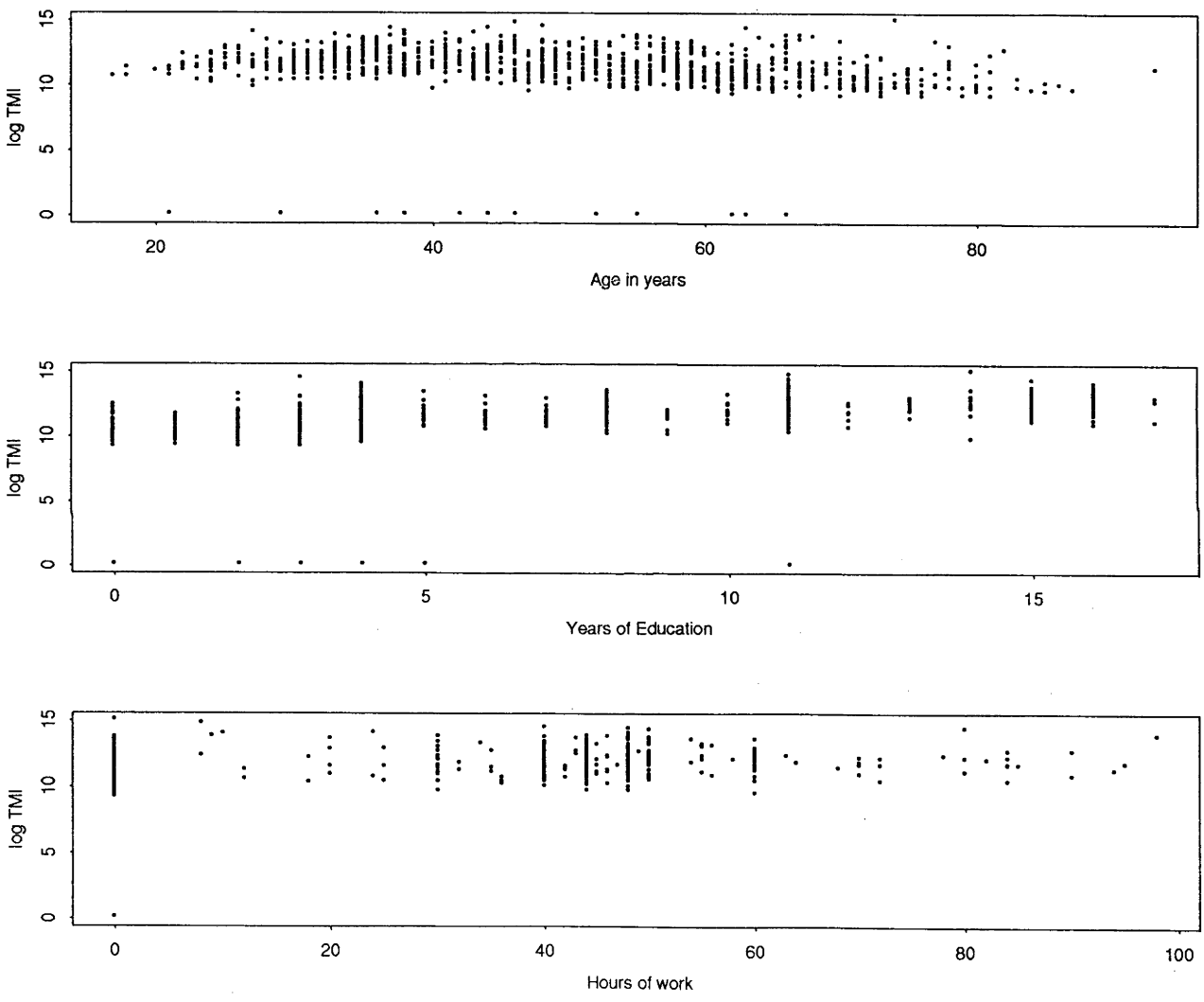


Figure 2.5 : Diagnostics for the Full Model by Robust regression with intercept

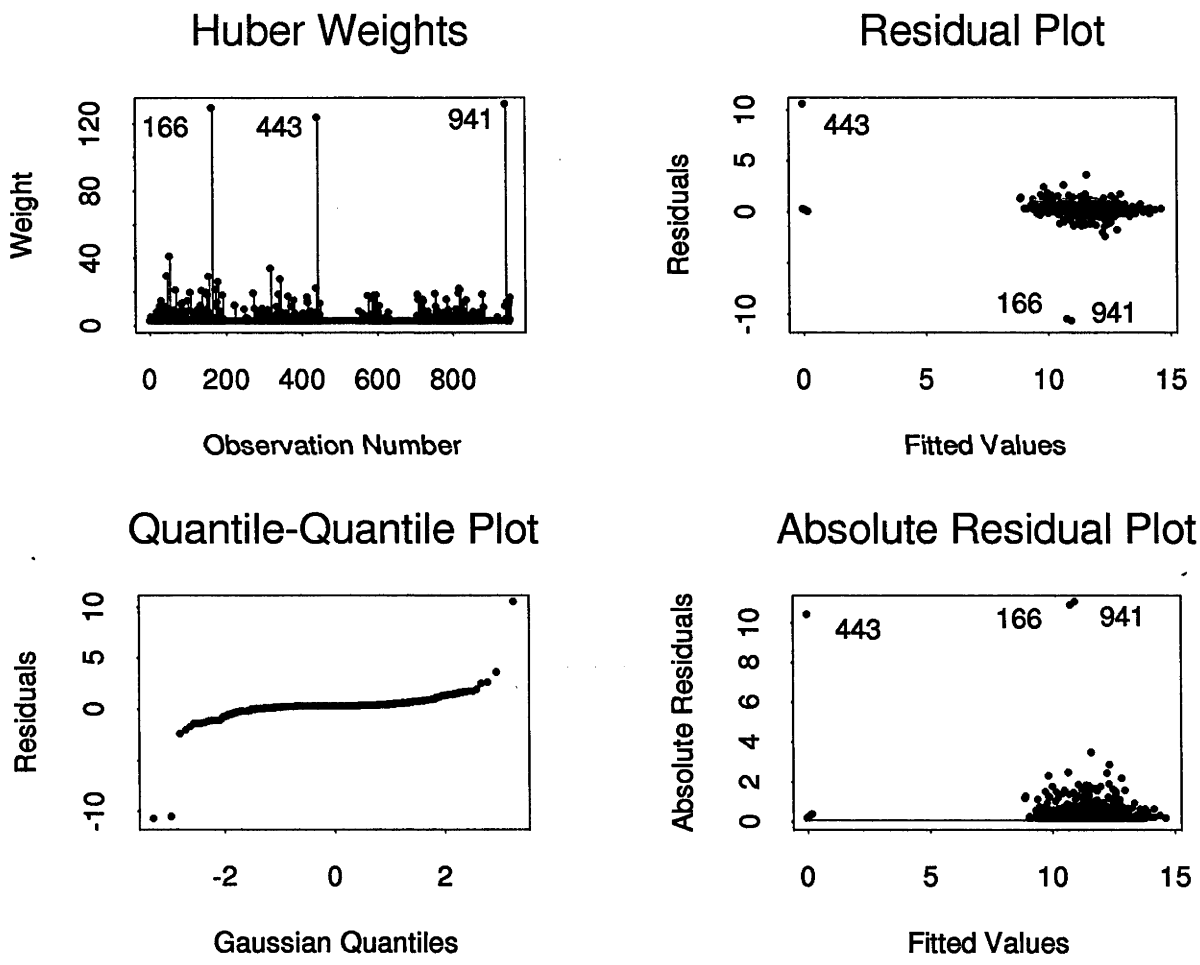


Figure 2.6 : Diagnostics for the Final Model by Robust regression with intercept. Two Predictors: log Proxy Income, Hourwork,

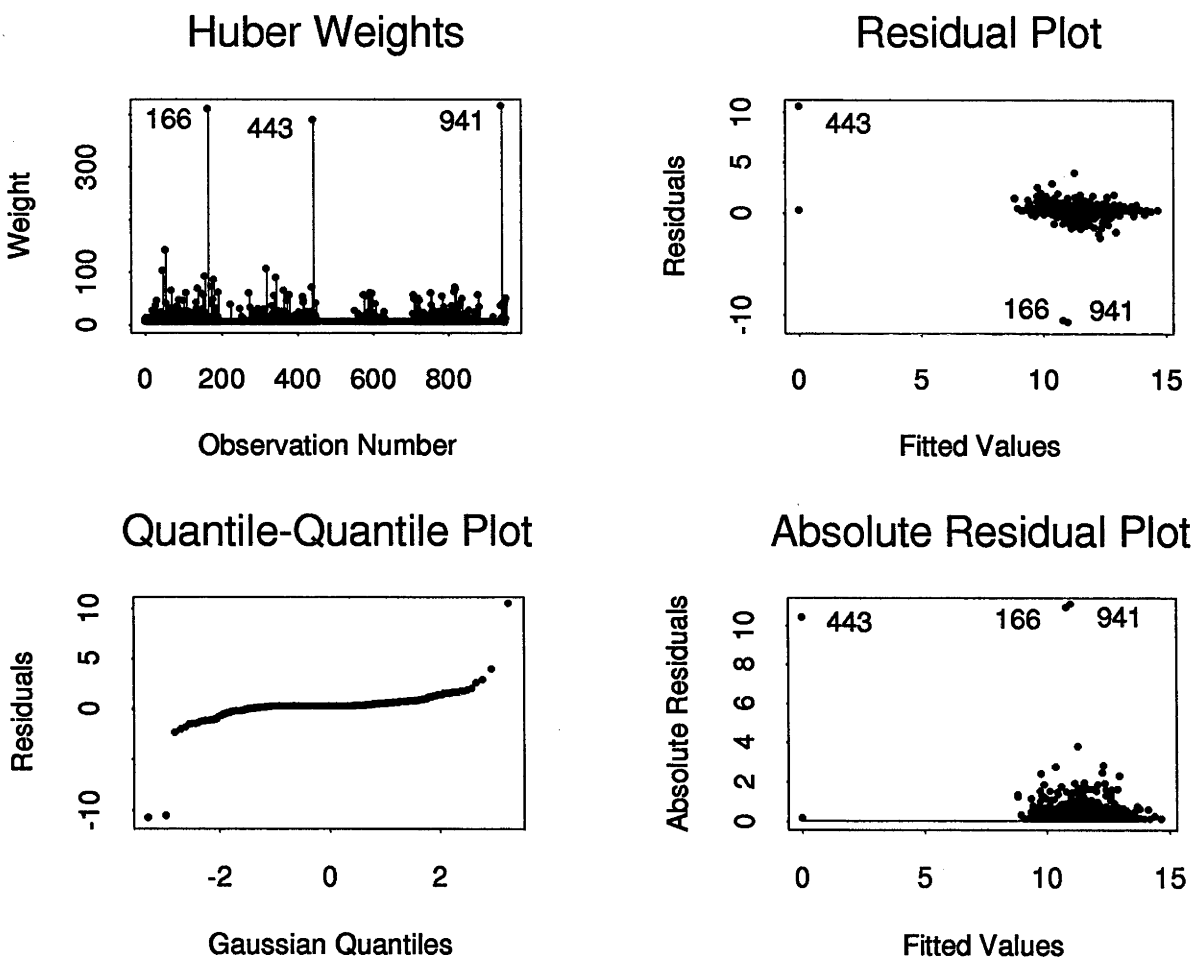


Figure 2.7 : Diagnostics for the Final Model by Robust Regression with intercept. 14 obs' excluded, Two Predictors : log PI, Hourwork,

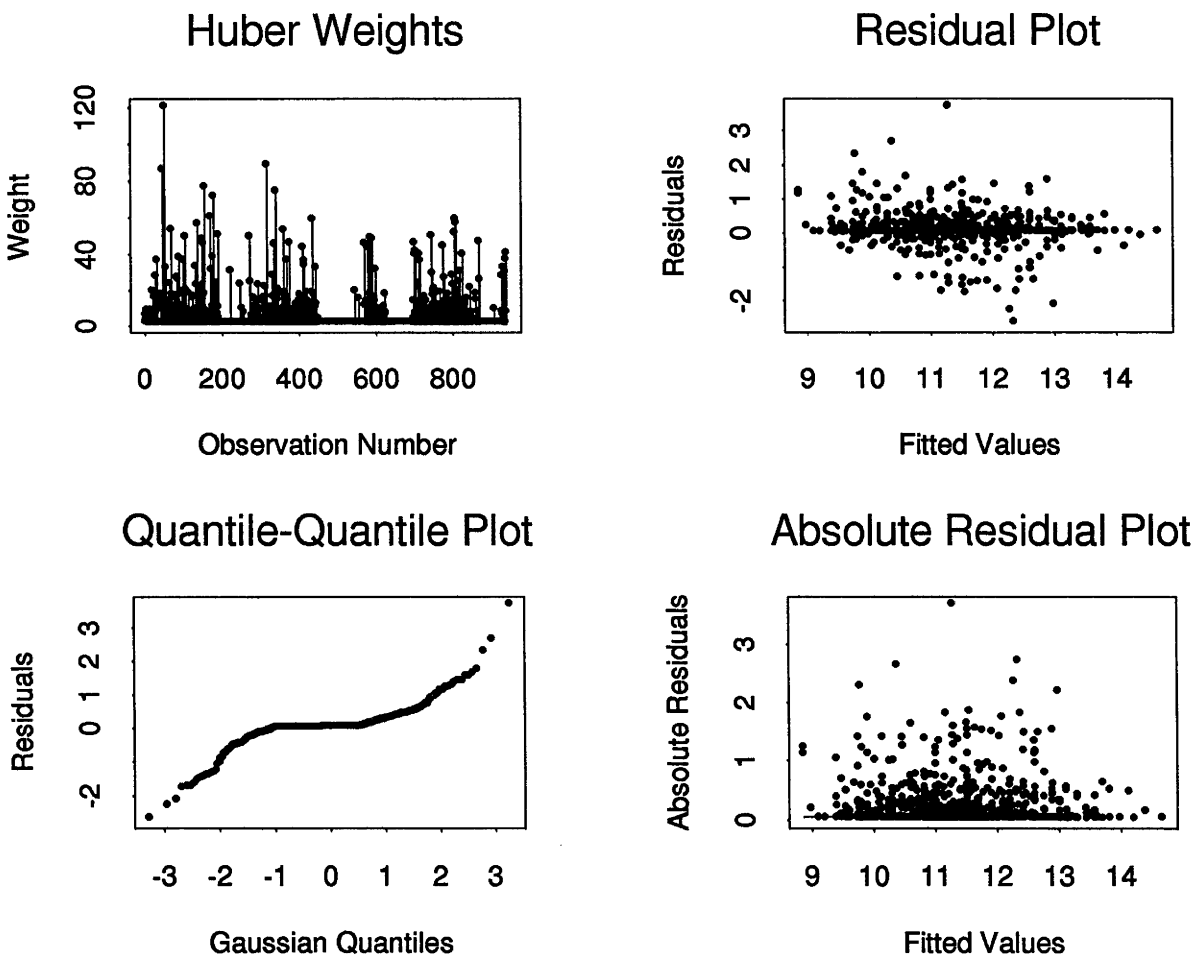


Figure 2.8 : Residual plot for the Final Model by Robust Regression with intercept. 14 obs' excluded, Predictors : log PI, Hourwork,

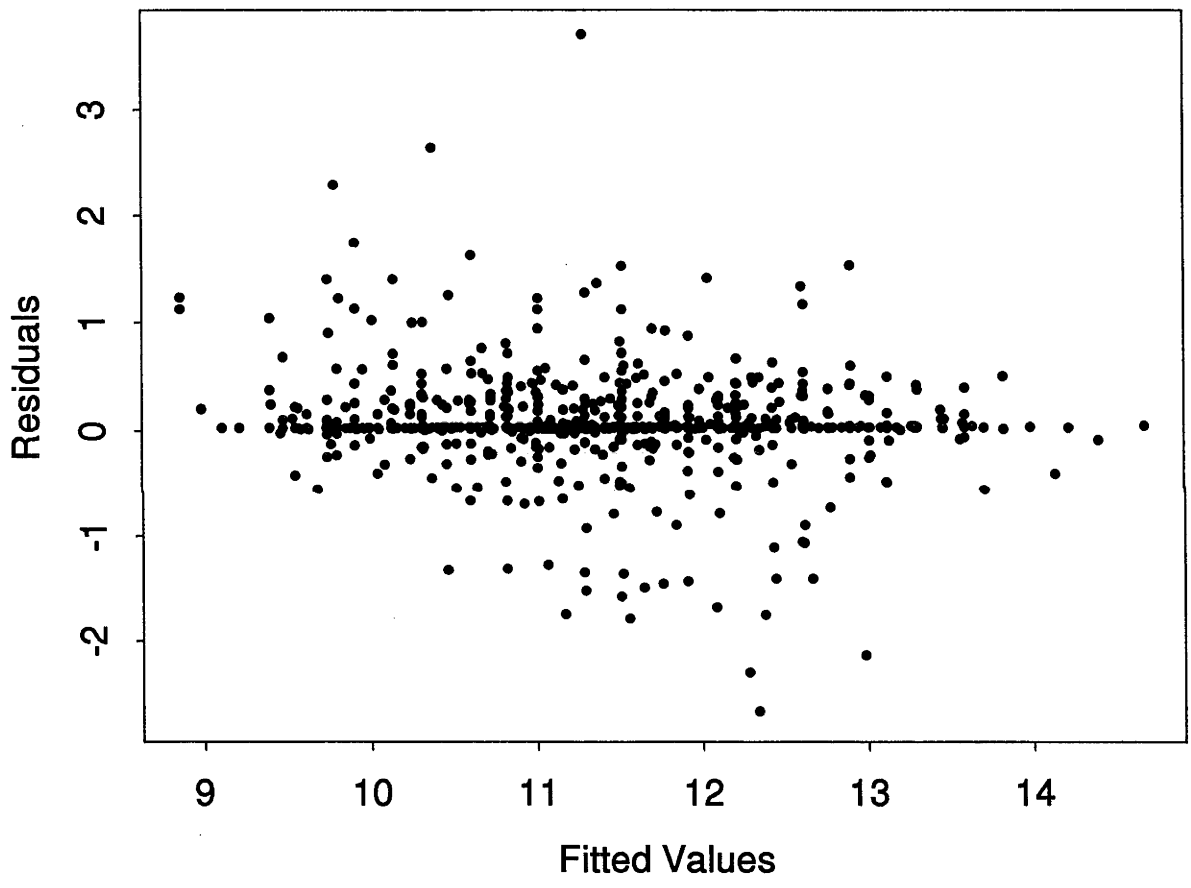


Figure 2.9 : Diagnostics for the Full Model (17 explanatory variables) without PI by Robust regression with intercept

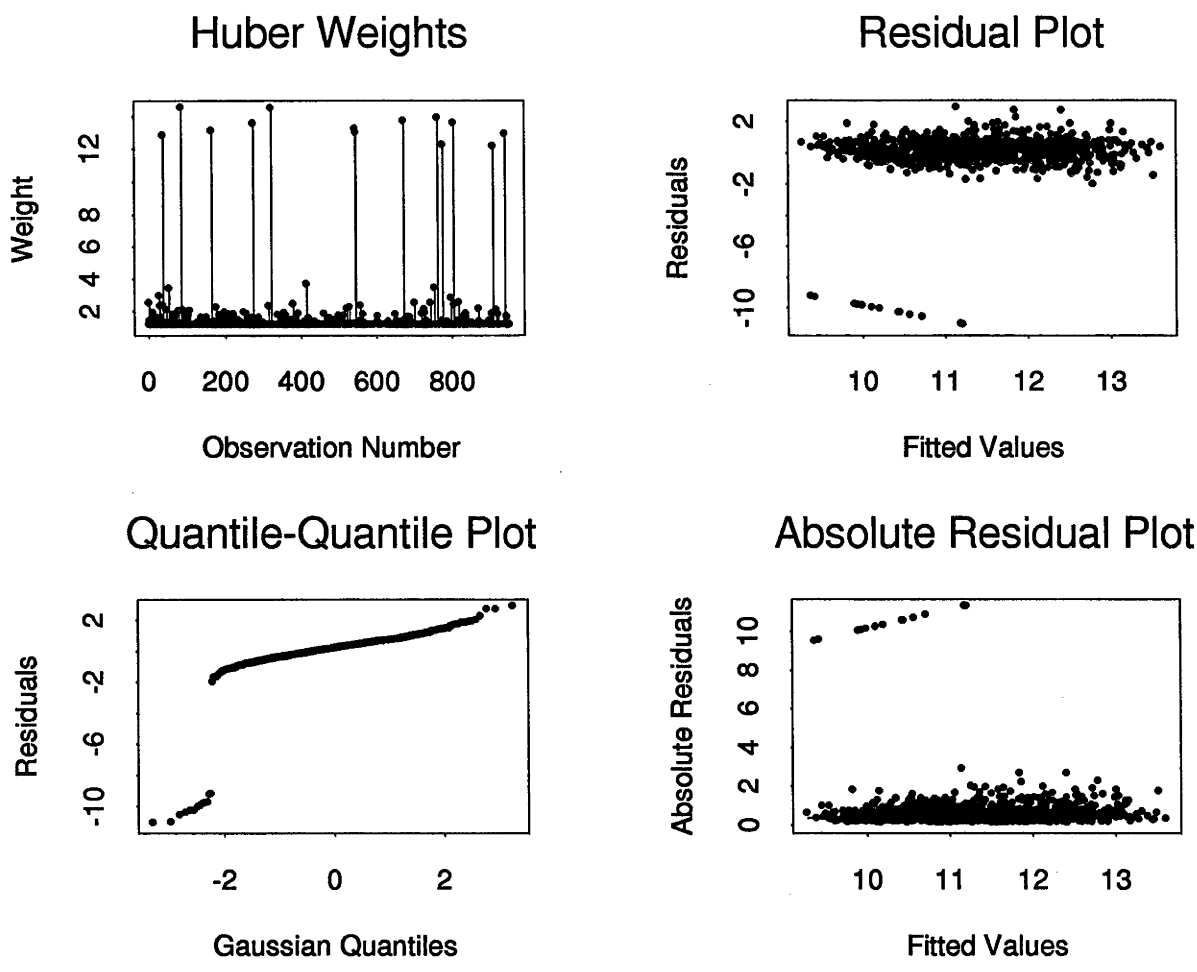
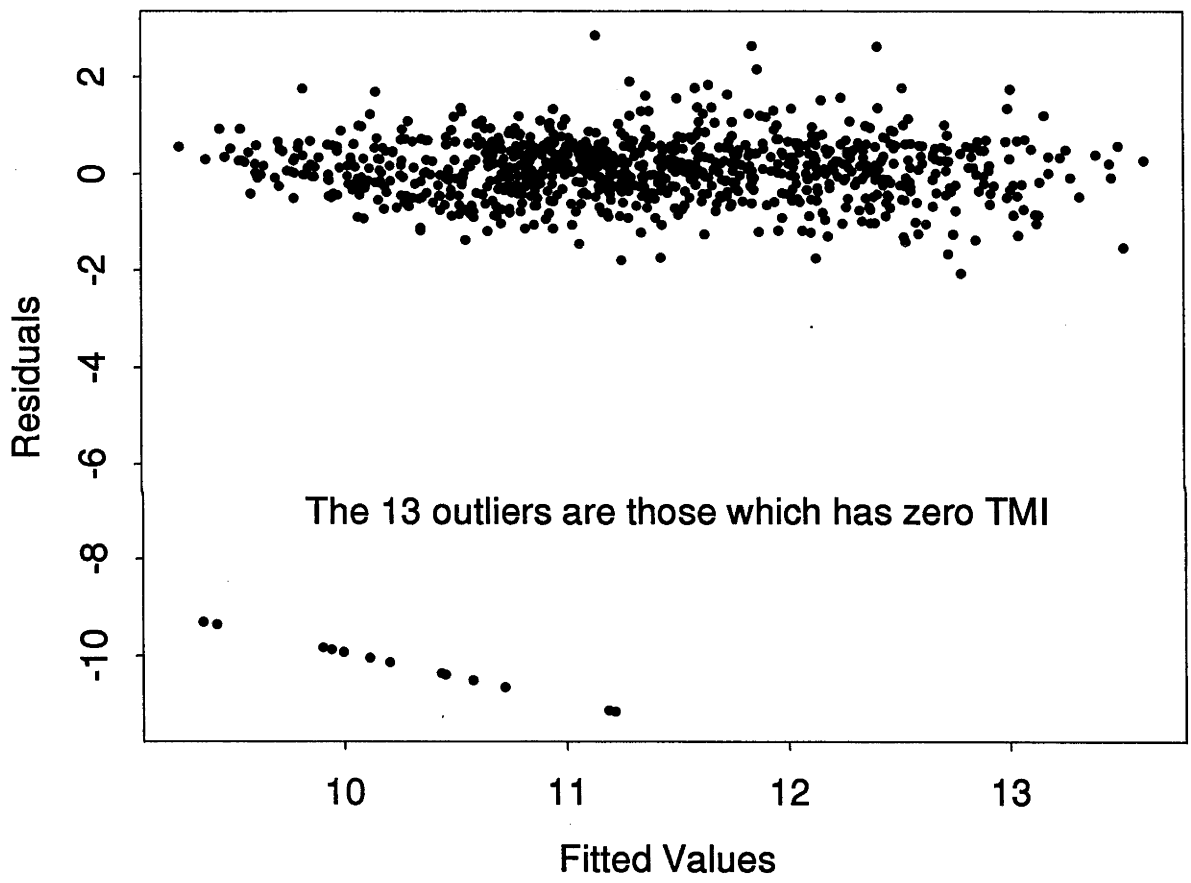


Figure 2.10 : Residual plot for the Full Model (17 explanatory variables) without PI by Robust Regression with intercept



3. Some Outlier Robust Estimators for the Population Total and the Population Distribution Function

The derivation of some estimators of the Population Total and the Population Distribution Function within a general linear superpopulation model framework are presented here. Results for these estimators in a single benchmark variable situation appeared in Chambers (1986), Chambers and Dunstan (1986) and Welsh and Ronchetti (1994).

Let Y denote the survey variable, with values $Y_1,\ldots,Y_N$ for the N Population elements. Without loss of generality, let $Y_1,\ldots,Y_n$ represent the sample observations so the nonsample portion of Y is $Y_{n+1},\ldots,Y_N$. Also suppose a multiple number of benchmark variables are available and are denoted $X_1,\ldots X_p$. Suppose that their values for the entire population are known. Then by assuming a linear model as the superpopulation model for Y, as in equation (2.1), we can write

$$\mathrm{Y} = \begin{pmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \end{pmatrix} \beta + e = \mathrm{X}\beta + e$$

where

$$\mathrm{Y}_1 = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathrm{Y}_2 = \begin{pmatrix} y_{n+1} \\ \vdots \\ y_N \end{pmatrix}$$

and

$$\mathrm{X}_1 = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \mathrm{X}_2 = \begin{pmatrix} x_{n+1,1} & \cdots & x_{n+1,p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix} = \begin{pmatrix} x_{n+1} \\ \vdots \\ x_N \end{pmatrix}$$

and suppose

$$\mathrm{var}(e) = \Sigma$$

such that

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad \Sigma_{11} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix} \qquad \Sigma_{22} = \begin{pmatrix} \sigma_{n+1,n+1} & \cdots & \sigma_{n+1,N} \\ \vdots & \ddots & \\ \sigma_{N,n+1} & \cdots & \sigma_{NN} \end{pmatrix}$$

and $\sqrt{\sigma_{ii}} = \sigma_i$ for $i = 1,\ldots,N$.

3.1 *Some Estimators of the Population Total*

Note that if T denotes the population total, then

$$T = \sum_{i=1}^{N} y_i = \sum_{i=1}^{n} y_i + \sum_{j=n+1}^{N} y_j = T_1 + T_2$$

say, where

$$T_1 = \sum_{i=1}^{n} y_i = 1_n' \mathrm{Y}_1 \qquad \text{and} \qquad T_2 = \sum_{j=n+1}^{N} y_j = 1_{N-n}' \mathrm{Y}_2$$

Once the sample has beeen observed, the values of $y_1,\ldots,y_n$ and hence T_1 are known. The problem then is to estimate the non-sample portion T_2 by the sample values $y_1,\ldots,y_n$. When there is no benchmark information, the classical design-based unbiased estimator is the usual number-raised estimator which is

$$\hat{T} = \frac{N}{n} \cdot 1_n' \mathrm{Y}_1 \tag{3.1}$$

When benchmark information is available and the population values of benchmark variables are $X_1,\ldots,X_p$, then under the superpopulation model $\mathrm{Y} = \mathrm{X}\beta + e$, the population total T can be estimated by the Least Squares method as

$$\hat{T}_{LS} = T_1 + \hat{T}_{2,LS}$$

where

$$\hat{T}_{2,LS} = 1'_{N-n}\mathrm{X}_2\hat{\beta}_{LS} \tag{3.2}$$

$$\hat{\beta}_{LS} = \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{Y}_1$$

This is the best linear unbiased estimator (BLUE) of T under the model (2.1) proposed by Brewer (1963) and Royall (1970).

However, it is well known that $\hat{\beta}_{LS}$ is sensitive to sample outliers especially and this sensitivity clearly carries over to $\hat{T}_{LS}$. An alternative approach is to replace $\hat{\beta}_{LS}$ by an outlier robust regression estimator such as the biweight estimator $\hat{\beta}_R$ of Beaton and Tukey (1974). So suppose we estimate T by

$$\hat{T}_R = T_1 + 1'_{N-n}\mathrm{X}_2\hat{\beta}_R \tag{3.3}$$

where $\hat{\beta}_R$ is the solution to

$$\mathrm{X}_1'\Sigma^{-1}\mathrm{M}\left(\mathrm{Y}_1 - \mathrm{X}_1\hat{\beta}_R\right) = 0$$

i.e.

$$\hat{\beta}_R = \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{M}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{M}\mathrm{Y}_1$$

where M is a diagonal matrix whose elements are

$$M_{(i,i)} = \eta\left(\frac{y_i - x_i\hat{\beta}_R}{c_E \cdot \hat{\sigma}_i}\right) \quad i = 1,\ldots,n$$

and

$$\eta(x) = \left(1 - x^2\right)^2 I(|x| \le 1)$$

Unfortunately, this $\hat{\beta}_R$ enables us to predict the bulk of the population very well but not the outliers and, in particular, not the outliers in the non-sample portions of the data. A natural thought would be to compromise between $\hat{\beta}_{LS}$ and $\hat{\beta}_R$. A simple way to achieve a compromise between the biweight and the least squares estimator is to modify the biweight fit by expanding it towards the least squares fit or vice versa. This was proposed by Chambers (1986) and called the bias-calibrated estimator. The derivation is as follows: Write

$$\begin{aligned}
\hat{T}_{LS} &= T_1 + 1'_{N-n}\mathrm{X}_2\hat{\beta}_{LS} \\
&= T_1 + 1'_{N-n}\left[\mathrm{X}_2\left(\theta + \hat{\beta}_{LS} - \theta\right)\right] \\
&= T_1 + 1'_{N-n}\mathrm{X}_2\theta + 1'_{N-n}\mathrm{X}_2\left(\hat{\beta}_{LS} - \theta\right)
\end{aligned}$$

and note that

$$\begin{aligned}
1'_{N-n}\mathrm{X}_2\left(\hat{\beta}_{LS} - \theta\right) &= 1'_{N-n}\mathrm{X}_2\left\{\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{Y}_1 - \theta\right\} \\
&= 1'_{N-n}\mathrm{X}_2\left\{\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{Y}_1 - \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)\theta\right\} \\
&= 1'_{N-n}\mathrm{X}_2\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}(\mathrm{Y}_1 - \mathrm{X}_1\theta)
\end{aligned}$$

So suppose b_n is an outlier robust estimator of β and let ψ be an appropriate function, then a robust estimator of T can be obtained by

$$\hat{T}_n = T_1 + 1'_{N-n}\mathrm{X}_2 b_n + 1'_{N-n}\mathrm{X}_2\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-\frac{1}{2}}\psi\left\{\Sigma_{11}^{-\frac{1}{2}}\left(\mathrm{Y}_1 - \mathrm{X}_1 b_n\right)\right\}$$

Hence if $c\Psi\left(\frac{x}{c}\right) = \max\left(-c, \min(x,c)\right)$, the Chambers (1986) bias-calibrated estimator is obtained as

$$\begin{aligned}
\hat{T}_C(c) &= T_1 + 1'_{N-n}\mathrm{X}_2\hat{\beta}_R + c\cdot 1'_{N-n}\mathrm{X}_2\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-\frac{1}{2}}\Psi\left\{c^{-1}\Sigma_{11}^{-\frac{1}{2}}\left(\mathrm{Y}_1 - \mathrm{X}_1\hat{\beta}_R\right)\right\} \\
&= T_1 + 1'_{N-n}\mathrm{X}_2\left\{\hat{\beta}_R + c\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-\frac{1}{2}}\Psi\left[c^{-1}\Sigma_{11}^{-\frac{1}{2}}\left(\mathrm{Y}_1 - \mathrm{X}_1\hat{\beta}_R\right)\right]\right\} \\
&= T_1 + 1'_{N-n}\mathrm{X}_2\hat{\beta}_C(c) \qquad (3.4)
\end{aligned}$$

where

$$\hat{\beta}_C(c) = \hat{\beta}_R + c\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-\frac{1}{2}}\Psi\left[c^{-1}\Sigma_{11}^{-\frac{1}{2}}\left(\mathrm{Y}_1 - \mathrm{X}_1\hat{\beta}_R\right)\right]$$

Since $\Psi(\cdot)$ is the Huber-psi function, it is obvious that when $c = 0$, $\hat{\beta}_C(c)$ is just the biweight estimator $\hat{\beta}_R$ and when c tends to infinity, $\hat{\beta}_C(c)$ becomes the least squares estimator because

$$\begin{aligned}
\hat{\beta}_C(c) &= \hat{\beta}_R + c\left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-\frac{1}{2}}\left[c^{-1}\Sigma_{11}^{-\frac{1}{2}}\left(\mathrm{Y}_1 - \mathrm{X}_1\hat{\beta}_R\right)\right] \\
&= \hat{\beta}_R + \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{Y}_1 - \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\hat{\beta}_R \\
&= \left(\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\Sigma_{11}^{-1}\mathrm{Y}_1 \\
&= \hat{\beta}_{LS}
\end{aligned}$$

3.2 *Decomposition of Population Distribution Function*

For a finite population, if $I(\cdot)$ is the usual indicator function then the population distribution function can be defined as

$$
\begin{aligned}
F(t) &= \frac{1}{N}\sum_{i=1}^{N} I(Y_i \le t), \qquad t \in R \\
&= \frac{1}{N}\left\{\sum_{i=1}^{n} I(Y_i \le t) + \sum_{i=n+1}^{N} I(Y_i \le t)\right\} \\
&= \frac{1}{N}\left\{n \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} I(Y_i \le t) + (N-n) \cdot \frac{1}{N-n} \cdot \sum_{i=n+1}^{N} I(Y_i \le t)\right\} \\
&= \frac{1}{N}\{nF_1(t) + (N-n)F_2(t)\} \\
&= \frac{n}{N}F_1(t) + \frac{N-n}{N}F_2(t)
\end{aligned}
$$

Therefore, the finite population distribution function can be decomposed into two, one is the sample distribution function and the other is the non-sample population distribution function. In any sample survey, values of $Y_1,\ldots,Y_n$ are known, so the problem is to estimate $F_2(t)$. That is

$$
\hat{F}(t) = \frac{n}{N}F_1(t) + \frac{N-n}{N}\hat{F}_2(t)
$$

where

$$
F_2(t) = \frac{1}{N-n}\sum_{i=n+1}^{N} I(Y_i \le t)
$$

3.3 *Some Distribution Function Estimators*

If benchmark informtion is not available, the simplest estimator of the distribution function of Y is just the sample distrbution function. i. e.

$$\hat{F}(t) = F_1(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le t) \tag{3.5}$$

but when benchmark information is available then under the superpopulation model $Y = X\beta + e$, a very simple way to estimate the distribution function (Welsh and Ronchetti (1994)) is to first construct the completed data sets $D = \left\{y_1, \ldots, y_n, x_{n+1}\hat{\beta}, \ldots, x_N\hat{\beta}\right\}$ then obtain $\hat{F}_2(t)$ as

$$\hat{F}_2(t) = \frac{1}{N-n}\sum_{i=n+1}^{N} I(x_i\hat{\beta} \le t)$$

If Least Squares estimate of β is used then $F_2(t)$ can be estimated as

$$\hat{F}_{2,LS}(t) = \frac{1}{N-n}\sum_{i=n+1}^{N} I(x_i\hat{\beta}_{LS} \le t) \tag{3.6}$$

But when the Biweight estimator of β is used, we obtain

$$\hat{F}_{2,R}(t) = \frac{1}{N-n}\sum_{i=n+1}^{N} I(x_i\hat{\beta}_{R} \le t) \tag{3.7}$$

If furthermore the Bias-Calibrated estimator of β is employed, then

$$\hat{F}_{2,C}(t,c) = \frac{1}{N-n}\sum_{i=n+1}^{N} I(x_i\hat{\beta}_{C}(c) \le t) \tag{3.8}$$

It is obvious that the four estimators from (3.5) to (3.8) described above are directly associated with the four estimators of population total of (3.1) to (3.4). Some other estimators of the distribution function are derived as the following.

Suppose Y and X are any two random variables and $h(\cdot)$ is an arbitrary function. If X is used to estimate Y, then

$$\begin{aligned} E\{Y-h(X)\}^2 &= \{E[Y-h(X)]\}^2 + V\{Y-h(X)\} \\ &= A^2 + V\{Y-E(Y|X)+E(Y|X)-h(X)\} \\ &= A^2 + V\{Y-E(Y|X)\}+V\{E(Y|X)-h(X)\}+2C \end{aligned}$$

where

$$A = E[Y-h(X)]$$
$$C = Cov\{Y-E(Y|X), E(Y|X)-h(X)\}$$

Since $E(Y|X)-h(X)$ is a function of X only, then by putting $d(X)=E(Y|X)-h(X)$, C can be written as

$$\begin{aligned} C &= Cov\{Y-E(Y|X), d(X)\} \\ &= E\{[Y-E(Y|X)]\cdot d(X)\}-E\{Y-E(Y|X)\}\cdot E\{d(X)\} \\ &= E\{E\{[Y-E(Y|X)]\cdot d(X)|X\}\} \\ &= E\{d(X)\cdot E\{[Y-E(Y|X)]\cdot|X\}\} \\ &= E\{d(X)\cdot[E(Y|X)-E(Y|X)]\} \\ &= 0 \end{aligned}$$

Therefore

$$\begin{aligned} E\{Y-h(X)\}^2 &= A^2 + V\{Y-E(Y|X)\}+V\{E(Y|X)-h(X)\} \\ &\geq V\{Y-E(Y|X)\} \end{aligned}$$

$$= E\{Y - E(Y|X)\}^2$$

Hence, $E(Y|X)$ is the best (minimum mean square error) predictor among all the other predictors based on the observed X. With this in mind, to find an estimator of $F_2(t)$ we look at the conditional expectation of $F_2(t)$ given the sample, say

$$\begin{aligned} EF_2(t) &= E\left\{\frac{1}{N-n}\sum_{j=n+1}^{N} I(y_j \le t)\right\} \\ &= E\left\{\frac{1}{N-n}\sum_{j=n+1}^{N} I\left(\frac{y_j - x_j\beta}{\sigma_j} \le \frac{t - x_j\beta}{\sigma_j}\right)\right\} \\ &= \frac{1}{N-n}\sum_{j=n+1}^{N} E\left\{I\left[\frac{y_j - x_j\beta}{\sigma_j} \le \frac{t - x_j\beta}{\sigma_j}\right]\right\} \\ &= \frac{1}{N-n}\sum_{j=n+1}^{N} G\left\{\frac{t - x_j\beta}{\sigma_j}\right\} \end{aligned}$$

where

$$G(r) = P\left\{\frac{y_j - x_j\beta}{\sigma_j} \le r\right\}$$

It is reasonable to estimate $G(r)$ by

$$\hat{G}(r) = \frac{1}{n}\sum_{i=1}^{n} I\left\{\frac{y_i - x_i\hat{\beta}_{LS}}{\hat{\sigma}_i} \le r\right\}$$

Motivated by this argument, we can rewrite $F_2(t)$ as

$$F_2(t) = \frac{1}{N-n}\sum_{j=n+1}^{N} I(y_j \le t) = \frac{1}{N-n}\sum_{j=n+1}^{N} I\left(\frac{y_j - x_j\beta}{\sigma_j} \le \frac{t - x_j\beta}{\sigma_j}\right)$$

and estimate $F_2(t)$ by

$$\hat{F}_{2,CD-ls}(t) = \frac{1}{N-n}\sum_{j=n+1}^{N}\frac{1}{n}\sum_{i=1}^{n} I\left\{\frac{y_i - x_i\hat{\beta}_{LS}}{\hat{\sigma}_i} \le \frac{t - x_j\hat{\beta}_{LS}}{\hat{\sigma}_j}\right\}$$

$$= \frac{1}{n(N-n)}\sum_{j=n+1}^{N}\sum_{i=1}^{n} I\left\{x_j\hat{\beta}_{LS} + \hat{\sigma}_j\left(\frac{y_i - x_i\hat{\beta}_{LS}}{\hat{\sigma}_i}\right) \le t\right\} \qquad (3.9)$$

This is the population distribution function estimator proposed by Chambers and Dunstan (1986) and it will be called as the Chambers and Dunstan least squares estimator $\hat{F}_{2,CD}(t)$ later on.

Welsh and Ronchetti (1994) suggested that since the finite population total T can be written as

$$T = N\int_{-\infty}^{\infty} t \cdot dF(t)$$

this means that the problem of estimating T can be subsumed within that of estimating F. Hence the non-sample total estimate can be derived from (3.4) and be written as

$$\hat{T}_{2,WR-ls} = 1'_{N-n}\mathrm{X}_2\hat{\beta}_{LS} + \frac{1}{n}\cdot tr\left(\Sigma_{22}^{\frac{1}{2}}\right)\cdot 1'_n\left\{\Sigma_{11}^{-\frac{1}{2}}\left(\mathrm{Y}_1 - \mathrm{X}_1\hat{\beta}_{LS}\right)\right\} \qquad (3.10)$$

and I shall call this the Welsh and Ronchetti least squares estimator.

Welsh and Ronchetti (1994) also suggested that if estimators other than $\hat{\beta}_{LS}$ in the Chambers and Dunstan least squares estimator is used, for example, the bias-calibrated estimator $\hat{\beta}_C(c)$, then $F_2(t)$ can be estimated as

$$\hat{F}_{2,WR-bc}(t)=\frac{1}{n(N-n)}\sum_{j=n+1}^{N}\sum_{i=1}^{n}I\left\{x_j\hat{\beta}_C(c)+\hat{\sigma}_j\left(\frac{y_i-x_i\hat{\beta}_C(c)}{\hat{\sigma}_i}\right)\le t\right\} \tag{3.11}$$

and this will be called as the Welsh and Ronchetti bias calibrated estimator of the population distribution function. By the same argument of (3.10), the non-sample total estimator corresponding to (3.10) is

$$\hat{T}_{2,WR-BC}(c)=1'_{N-n}\mathrm{X}_2\hat{\beta}_C(c)+\frac{1}{n}\cdot tr\left(\Sigma_{22}^{\frac{1}{2}}\right)\cdot 1'_n\left\{\Sigma_{11}^{-\frac{1}{2}}\left(\mathrm{Y}_1-\mathrm{X}_1\hat{\beta}_C(c)\right)\right\} \tag{3.12}$$

It shall be called as the Welsh and Ronchetti bias calibrated estimator.

Furthermore, if the Biweight estimator $\hat{\beta}_R$ is used instead of $\hat{\beta}_{LS}$ nor $\hat{\beta}_C(c)$ in (3.9), then $F_2(t)$ can be estimated as

$$\hat{F}_{2,WR-bi}(t)=\frac{1}{n(N-n)}\sum_{j=n+1}^{N}\sum_{i=1}^{n}I\left\{x_j\hat{\beta}_R+\hat{\sigma}_j\left(\frac{y_i-x_i\hat{\beta}_R}{\hat{\sigma}_i}\right)\le t\right\} \tag{3.13}$$

It will be denoted Welsh and Ronchetti biweight estimator of population distribution function and the corresponding non-sample total estimator is

$$\hat{T}_{2,WR-bi}=1'_{N-n}\mathrm{X}_2\hat{\beta}_R+\frac{1}{n}\cdot tr\left(\Sigma_{22}^{\frac{1}{2}}\right)\cdot 1'_n\left\{\Sigma_{11}^{-\frac{1}{2}}\left(\mathrm{Y}_1-\mathrm{X}_1\hat{\beta}_R\right)\right\} \tag{3.14}$$

This willl be called as the Welsh and Ronchetti biweight estimator of the non-sample total. Note that this is different from the $\hat{T}_{2,R}=1'_{N-n}\mathrm{X}_2\hat{\beta}_R$ in (3.3) and this $\hat{T}_{2,WR-Bi}$ is just the $\hat{T}_{2,WR-BC}(c)$ when $c=0$. When c tends to infinity, $\hat{T}_{2,WR-BC}(c)$ becomes $\hat{T}_{2,WR-LS}$.

Welsh and Ronchetti (1994) also suggested a sophisticated alternative to the Chambers and Dunstan (1986) estimator which used the biweight fit together with the bounded residuals

$$\hat{F}_{2,WR-r}(t,c)=\frac{1}{n(N-n)}\sum_{j=n+1}^{N}\sum_{i=1}^{n}I\left\{x_j\hat{\beta}_R+c\cdot\hat{\sigma}_j\cdot\Psi\left[\frac{y_i-x_i\hat{\beta}_R}{c\cdot\hat{\sigma}_i}\right]\le t\right\} \tag{3.15}$$

This will be called the Welsh and Ronchetti robust estimator of the population distribution and note that the corresponding non-sample estimate is

$$\hat{T}_{2,WR-r}(c)=1'_{N-n}\mathrm{X}_2\hat{\beta}_R+\frac{1}{n}\cdot tr\left(\Sigma_{22}^{\frac{1}{2}}\right)\cdot c\cdot 1'_n\Psi\left\{c^{-1}\cdot\Sigma_{11}^{-\frac{1}{2}}\left(\mathrm{Y}_1-\mathrm{X}_1\hat{\beta}_R\right)\right\} \tag{3.16}$$

Chapters 4 and 5 would investigate the performance of these estimators when applied to the Brazilian data.

4. Estimating the Population Total of the Brazilian Data

In this chapter the performance of each population total estimator applied to the Brazilian data is compared and a discussion of the role of bias calibration in outlier robust estimation of population total is given.

4.1 *Various Estimators under the Superpopulation Model of the Brazilian Data*

The various estimators of population total described in chapter 3 are under a general linear model framework and therefore when being applied to the Brazilian data, some adjustments have to be made. Under the superpopulation model of the Brazilian data, by (2.3) and (2.4) the variance-covariance structure of e is just $\sigma^2 I$. where a robust estimate of σ is

$$\hat{\sigma} = \frac{median\left|\left(y_i - x_i\hat{\beta}_R\right) - median\left(y_i - x_i\hat{\beta}_R\right)\right|}{0.6745} \qquad i = 1,\ldots,n$$

Hence the population total estimators mentioned in chapter 3 can be simplified as the following:

In general, $\hat{T} = T_1 + \hat{T}_2 = T_1 + 1'_{N-n}\mathrm{X}_2\hat{\beta}$

(1) Number-raised estimator

$$\hat{T} = \frac{N}{n}\cdot 1'_n \mathrm{Y}_1 \tag{4.1}$$

(2) Brewer (1963) and Royall (1970) 's least squares estimator (3.2)

$$\hat{\beta}_{LS} = \left(\mathrm{X}_1'\mathrm{X}_1\right)^{-1}\mathrm{X}_1'\mathrm{Y}_1 \tag{4.2}$$

(3) Beaton and Tukey (1974) 's biweight estimator (3.3)

$$\hat{\beta}_R = \left(X_1' M X_1\right)^{-1} X_1' M Y_1 \tag{4.3}$$

where M is defined as in (3.3).

(4) Chambers bias-calibrated estimator (3.4)

$$\hat{\beta}_C(c) = \hat{\beta}_R + c \cdot \hat{\sigma} \cdot \left(X_1' X_1\right)^{-1} X_1' \Psi\left\{\frac{Y_1 - X_1 \hat{\beta}_R}{c \cdot \hat{\sigma}}\right\} \tag{4.4}$$

Note that when $c = 0$, $\hat{\beta}_C(c) = \hat{\beta}_R$ and when $c = \infty$, $\hat{\beta}_C(c) = \hat{\beta}_{LS}$

(5) Welsh and Ronchetti least squares estimator (3.10)

$$\hat{T}_{2,WR-ls} = 1'_{N-n} X_2 \hat{\beta}_{LS} + \frac{N-n}{n} \cdot 1'_n \left\{Y_1 - X_1 \hat{\beta}_{LS}\right\} \tag{4.5}$$

(6) Welsh and Ronchetti biweight estimator (3.14)

$$\hat{T}_{2,WR-bi} = 1'_{N-n} X_2 \hat{\beta}_R + \frac{N-n}{n} \cdot 1'_n \left\{Y_1 - X_1 \hat{\beta}_R\right\} \tag{4.6}$$

(7) Welsh and Ronchetti bias calibrated estimator (3.12)

$$\hat{T}_{2,WR-bc}(c) = 1'_{N-n} X_2 \hat{\beta}_C(c) + \frac{N-n}{n} \cdot 1'_n \left\{Y_1 - X_1 \hat{\beta}_C(c)\right\} \tag{4.7}$$

Similar to (4.4), when $c = 0$, $\hat{T}_{2,WR-bc}(c) = \hat{T}_{2,WR-bi}$ and

when $c = \infty$, $\hat{T}_{2,WR-bc}(c) = \hat{T}_{2,WR-LS}$

(8) Welsh and Ronchetti robust estimator (3.16)

$$\hat{T}_{2,WR-r}(c) = 1'_{N-n} X_2 \hat{\beta}_R + \frac{N-n}{n} \cdot c \cdot \hat{\sigma} \cdot 1'_n \Psi\left\{\frac{Y_1 - X_1 \hat{\beta}_R}{c \cdot \hat{\sigma}}\right\} \tag{4.8}$$

4.2 *The Selected Simple Random Samples*

In order to investigate the performance of all the estimators and the role of bias calibration in outlier robust estimation described in chapter 3, three different yet representative samples of each one hundred households are selected from the population by simple random sampling. The characteristics of these three samples are such that they roughly represent three comprehensive categories of samples. That is, any simple random sample drawn from the population will inevitably fall into one of these three categories. Figure 4.1 shows the scatter plots (i.e. log TMI vs log PI) of these three samples together with the scatter plot of the population.

Sample 1 represents samples which do not contain outliers. Sample 2 represents samples without extreme outliers but a few mild negative outliers. Sample 3 represents samples with a few mild but not extreme positive outliers. That is

Sample 1 : no outliers
Sample 2 : sample with mild negative outliers
Sample 3 : sample with mild positive outliers

Recall from section (2.2) the exploratory data analysis showed that there are three extreme outliers out of the 954 households. Observations 166 and 941 are households with zero Total Monthly Income (TMI) but non-zero Proxy Income (PI) and household 443 has non-zero TMI but zero PI. As these three observations are so different from the rest of households and the chance of including any one in a sample of 100 out of 954 is not very high, they are not included in the three selected samples.

When the population total is estimated from the three samples under the Superpopulation Model II of (2.4), the three samples are no longer representative samples of the three categories mentioned previously because this model exlude Proxy Income as an explanatory variable. Under this model (with eight predictors), there is no simple way to describe the samples. Nevertheless, under the Model II, these three samples should be able to shed us some light on the performance of the various estimators and the role of the bias calibration in outlier robust estimation.

4.3 *Performance of the various Population Total Estimators on the Brazilian Data*

In this section, estimates of the population total from the three selected samples according to various estimators under Superpopulation Model I (2.3) and Superpopulation Model II (2.4) are presented. Note that the optimal c of the bias calibrated estimators is the value of c such that the bias calibrated estimate is the closest to the true total.

Table 4.1 : Estimates of Population Total from Sample 1 (no outliers)

Various Estimators	Model I	Model II
True Population Total	**146,522,571**	**146,522,571**
Number-raised	133,143,412	**133,143,412**
Least squares	137,183,110	127,473,871
Welsh and Ronchetti least squares	137,184,028	**127,475,180**
Biweight	134,905,432	120,901,780
Welsh and Ronchetti biweight	134,906,373	120,902,736
Chambers bias calibrated (optimal c)	138,123,891 c=1128	127,473,871 $c \geq 21$
Welsh and Ronchetti bias calibrated (optimal c)	**138,124,801 c=1128**	**127,475,180 $c \geq 21$**
Welsh and Ronchetti robust (optimal c)	134,906,374 $c \geq 1357$	120,902,739 $c \geq 2$

The residual plot of biweight fit under Model I (Figure 4.2) clearly shows that Sample 1 does not contain outliers but there are possibly three high influential observations at the bottom. However, when the same sample is analysed under Model II there are two extreme

outliers (Figure 4.3) and they are identified to be the households with zero TMI and zero PI. Thus it seems that Model II fails to account for households of this type. As a result, all estimators under both models tend to underestimate the true total.

Under Model I, all estimates are very close to the true total where the usual number-raised estimator is outperformed by all other estimators and among these model-based estimators, bias calibrated type estimators are able to yield the best estimates. Under Model II the number-raised estimate is the closest to the true value and the least squares type estimators are the best among all the model-based estimators.

Figures 4.2 and 4.3 show the effects of the choice of c on bias calibrated estimates of population total. Clearly, the effect of c on total estimates (c rises from zero to infinity) is not monotonic as one might think. This is a very interesting behaviour because when both the least Squares type estimators and the biweight type estimators tend to underestimate the true total, an appropriate choice of c of bias calibrated type estimators can yield an estimate which outperforms the estimates by least squares and biweight estimators. Apart from this, the range of Welsh and Ronchetti robust estimates influenced by the choice of c is surprisingly narrow and this can be very positive as well as negative because if the bias is large then the choice of c will have little effect of improving the estimate and if the bias is small, the estimate yielded by an inappropriate choice of c would still be close to the true value.

Results of population total estimates from Sample 2 are

Table 4.2 : Estimates of Population Total from Sample 2(negative outliers)

Various Estimators	Model I	Model II
True Population Total	**146,522,571**	**146,522,571**
Number-raised	118,681,700	118,681,700
Least squares	100,943,764	108,593,933
Welsh and Ronchetti least squares	100,944,726	108,594,972
Biweight	132,627,539	107,368,590
Welsh and Ronchetti biweight	132,628,369	107,369,641
Chambers bias calibrated (optimal c)	132,627,539 c=0	108,747,868 c=3
Welsh and Ronchetti bias calibrated (optimal c)	132,628,369 c=0	**108,748,903 c=3**
Welsh and Ronchetti robust (optimal c)	**132,628,394 c=10**	107,369,642 $c \geq 3$

In the case of Sample 2, all estimates under Model I are greatly affected by those negative outliers (Figure 4.4). Hence least squares type estimates are the farthest from the true value and biweight type estimates the closest. The Welsh and Ronchetti robust estimate is basically as good as and even better (with appropriate choice of c) than the biweight type estimators.

Whereas under Model II, least squares type estimates are the better and the best option is the Welsh and Ronchetti bias calibrated estimator. The situation here is slightly strange since the residual plot of the biweight fit (Figure 4.5) shows nothing peculiar and one would think this is an ideal residual plot. Yet all the estimates based on this sample are even worse than the estimates given by Sample 1 under Model II. Since Model II involves eight explanatory variables there is no simple explanation of why this occurs. But we can deduce from this that Model I is a better model than Model II. This is particularly true since we know the existence of Proxy Income (highly correlated with Total Monthly Income) as a predictor under Model I. Again it happens that with an appropriate choice of c, the Welsh and Ronchetti bias calibrated estimator surpasses the least squares type estimators. Of all the choices of c, range of the Welsh and Ronchetti robust estimates is again very narrow when compare with the other two bias calibrated estimators.

Table 4.3 : Estimates of Population Total from Sample 3 (positive outliers)

Various Estimators	Model I	Model II
True Population Total	**146,522,571**	**146,522,571**
Number-raised	204,786,877	204,786,877
Least squares	175,142,941	180,836,468
Welsh and Ronchetti least squares	175,144,211	180,837,701
Biweight	143,158,645	**151,264,527**
Welsh and Ronchetti biweight	143,160,395	151,265,942
Chambers bias calibrated (optimal c)	**147,092,490 c=3**	**151,264,527 c=0**
Welsh and Ronchetti bias calibrated (optimal c)	147,094,185 c=3	151,265,942 c=0
Welsh and Ronchetti robust (optimal c)	143,160,395 $c \geq 90$	151,265,381 c=0

In the situation of Sample 3, least squares type estimators under Model I overestimate the true value while biweight type estimators underestimate it. Thus the optimal type of estimators is the bias calibrated estimators which can even be made to hit the true total. Note that this situation of Sample 3 under Model I is analogous to the situation of beef farm samples discussed in Welsh and Ronchetti (1994) so that for samples with a relatively few representative positive outliers, bias calibration estimator is able to give a very accurate estimate with an appropriate choice of c. Under Model II, number raised estimator and least squares type estimators greatly overestimate the true total and the biweight type estimates are the closest to the true value.

Figure 4.6 and Figure 4.7 show that when c increases, bias calibrated type estimates under both models are also monotonic increasing.

4.4 *Discussion*

In general, for a sample without outliers (represented by Sample 1) from a linear superpopulation model, least squares estimators are the best solution in estimating the population total. For a sample containing negative outliers (represented by Sample 2), since the biweight fit in practice always tends to underestimate the true value, all estimators described so far underestimate the true population total and the biweight type estimators are preferred as the least squares fit is so sensitive to outliers. In such a case the bias calibrated estimators essentially do not offer much help as they generally yield an estimate lying between the biweight estimate and the least squares estimate. However, when it comes to the situation of sample containing positive outliers (represented by Sample 3), that is when least squares overestimates and the biweight underestimates, bias calibrated estimators can be made to yield the best estimate.

Based on a comparison of the various estimates of the three samples under both superpopulation Model I and II, we see that when the sample does not contain outliers, model based estimators are as good as the number raised estimator. However, when the sample contains outliers, number raised estimator can be even worse than the least squares estimate. Furthermore, Figure 4.5 has illustrated that a good diagnostic of a certain fit does not

guarantee a good estimate. More research is needed to tell us when a model based estimator of the population total is better than the number raised estimator for this situation.

It is also clear from the results of section 4.3 that the adjustments made from least squares to Welsh and Ronchetti least squares, biweight to Welsh and Ronchetti biweight and Chambers bias calibrated to Welsh and Ronchetti bias calibrated affect the estimates only very slightly. However, this empirical study reveals that the Welsh and Ronchetti robust estimator has more bias but less variance. It also has a tendency to underestimate the true value (probabily inherited from the biweight estimator) but in the case of a sample containing positive outliers, it would be a very good and attractive estimator because the optimal value of c is not known and a choice of a non-optimal c for the Welsh and Ronchetti robust estimator can still produce a very good estimate.

The non-monotonicity (Figure 4.2, Figure 4.3 and Figure 4.5) of bias calibrated estimators from biweight to least squares fit or vice versa is interesting. It happens in Sample 1 and Sample 2 (underestimation occurs in these two samples) that the bias calibrated estimators can yield an estimate not between the biweight and least squares estimates. This means in certain circumstances, such as when both least squares and biweight estimators underestimate the true value, the bias calibrated estimator is able to surpass their performance.

Another interesting point is, for the three bias calibrated estimators under both models, the choices of the optimal c are quite similar. More research may be needed to see if the value of the optimal c is model invariant.

4.5 *A Strategy of Estimating the Population Total*

I. When the ranom sample does not contain outliers, number raised estimator and model based estimators are compatible. However, if we are confident about the superpopulation model then the least squares estimator is the best option.

II. When the random sample contains negative outliers, a robust estimator such as the biweight estimator of a resonable superpopulation model surpasses all other estimators described in chapter 3. However, the results of section 4.3 illustrates that there is still room for

improvement of the biweight estimator in this situation. More research is needed to deal with this.

III. When the random sample contains positive outliers, bias calibrated estimators are vital in estimating the population total. However, the exact choice of c depends on the context and the nature of the outliers. For situation similar to the Brazilian data where the sample does not contain serious positive outliers but the mild one, the Welsh and Ronchetti robust estimator seems to be the best of all for its boundedness property for whatever the choice of c.

Figure 4.1 : Scatter Plots of log Total Monthly Income vs log Proxy Income for Population and the three Samples

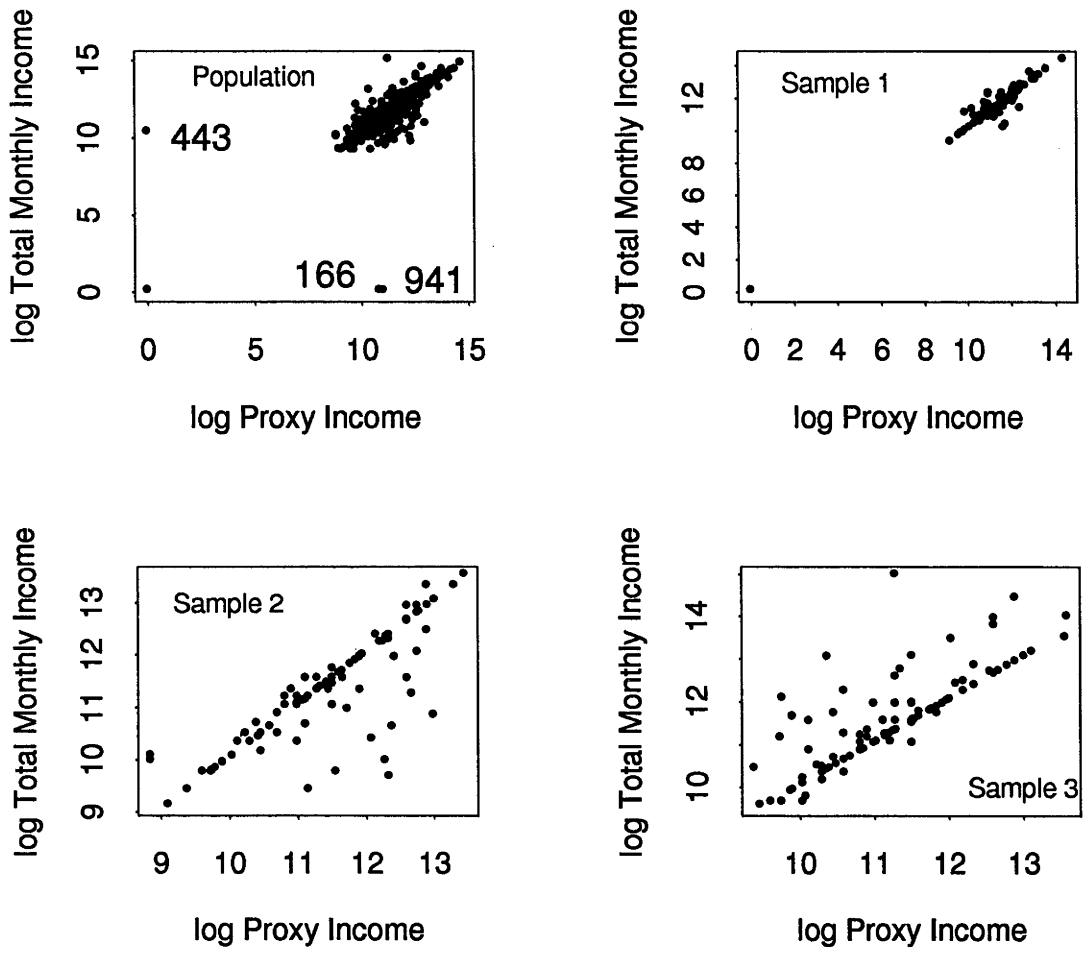
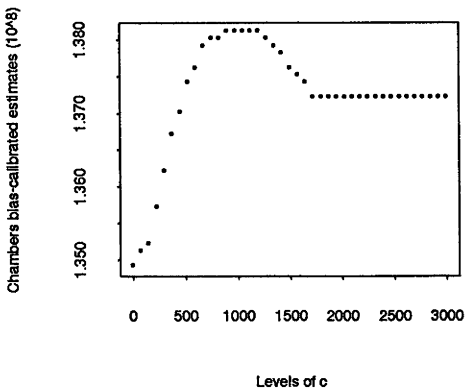
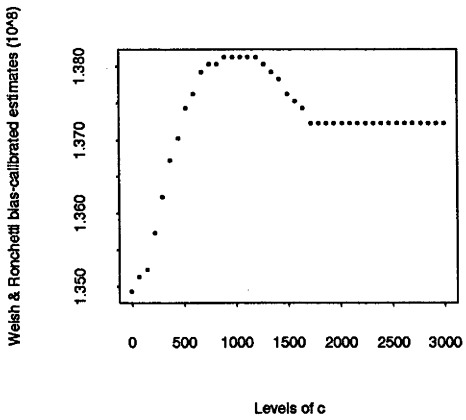
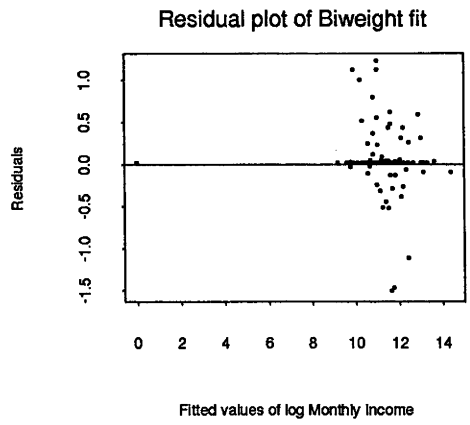


Figure 4.2 : Bias calibrated estimates of Population Total

Sample 1 under Model I







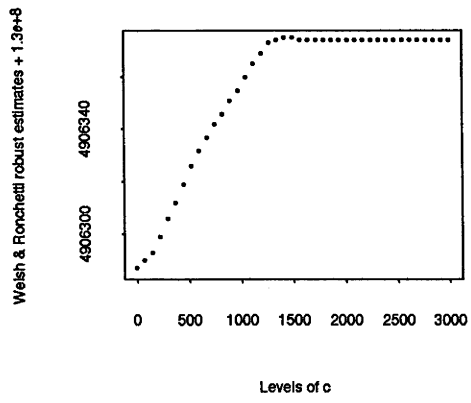


Figure 4.3 : Bias calibrated estimates of Population Total

Sample 1 under Model II

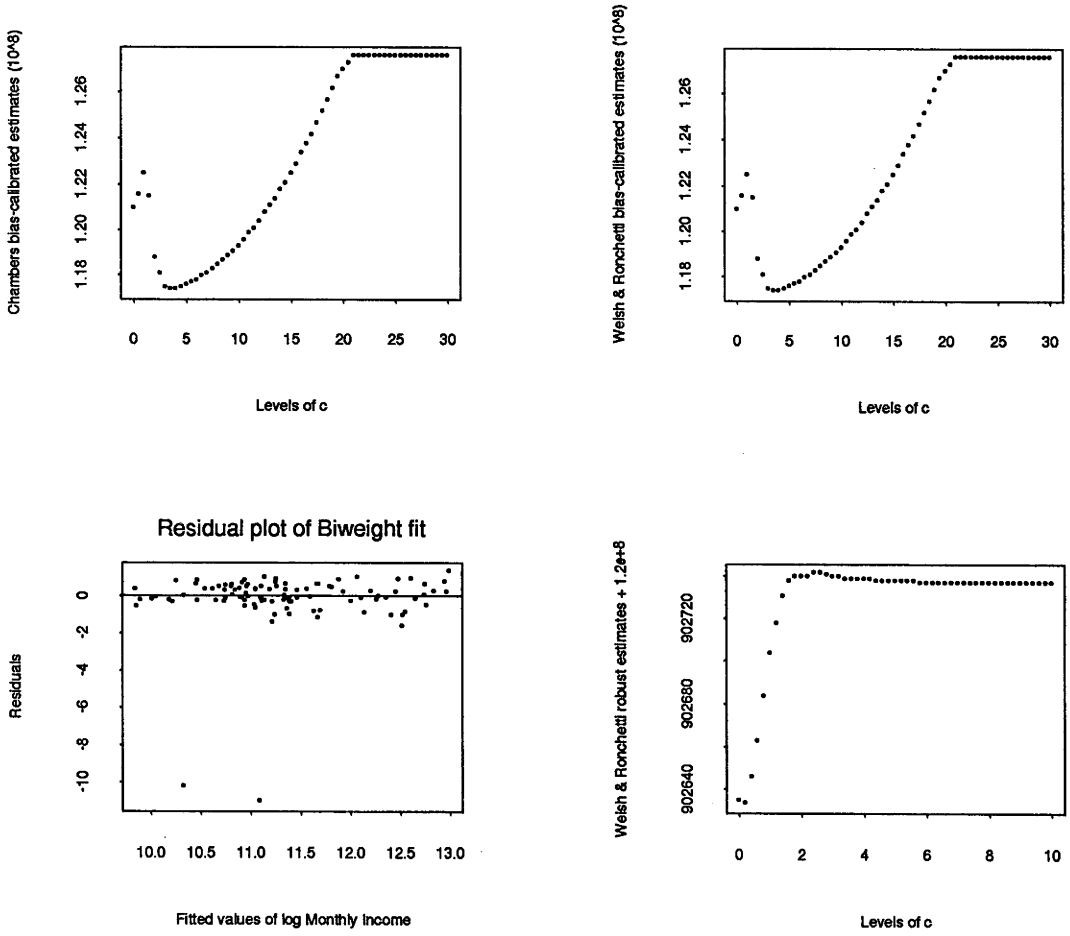


Figure 4.4 : Bias calibrated estimates of Population Total

Sample 2 under Model I

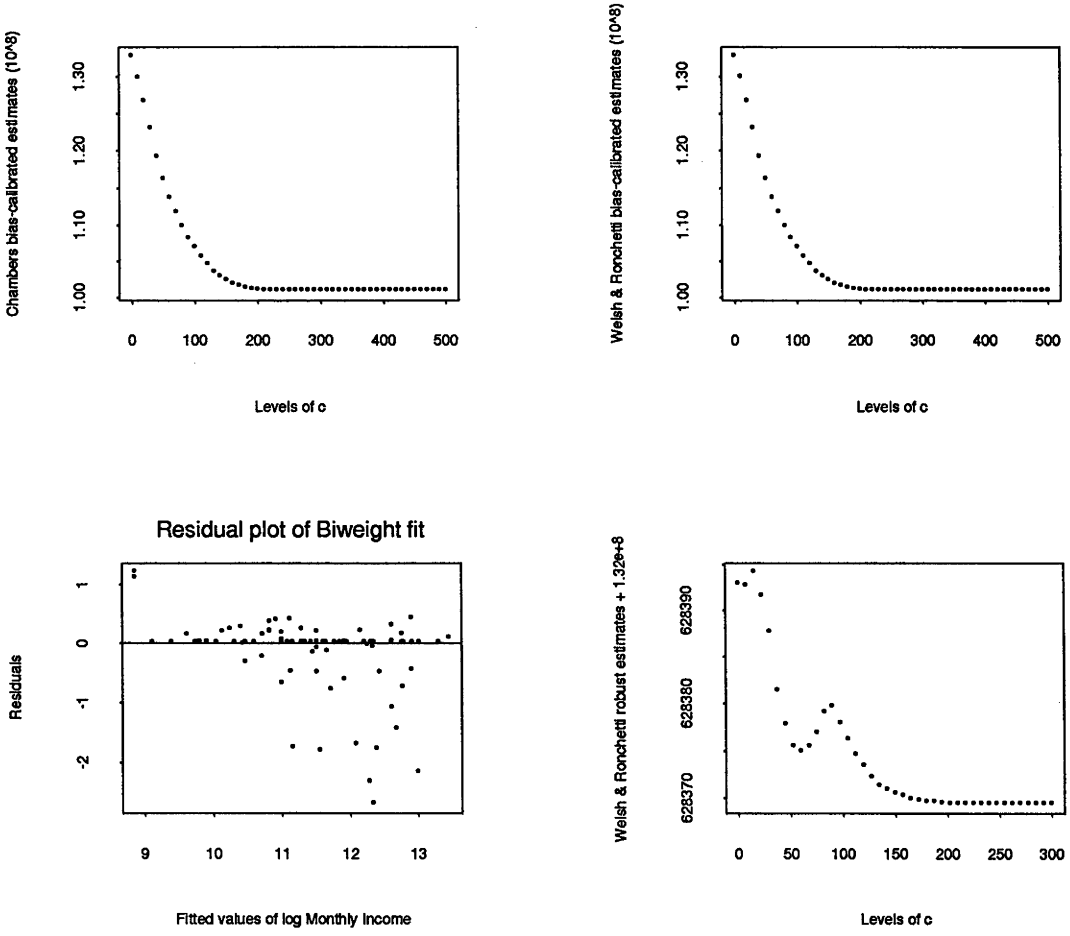


Figure 4.5 : Bias calibrated estimates of Population Total

Sample 2 under Model II

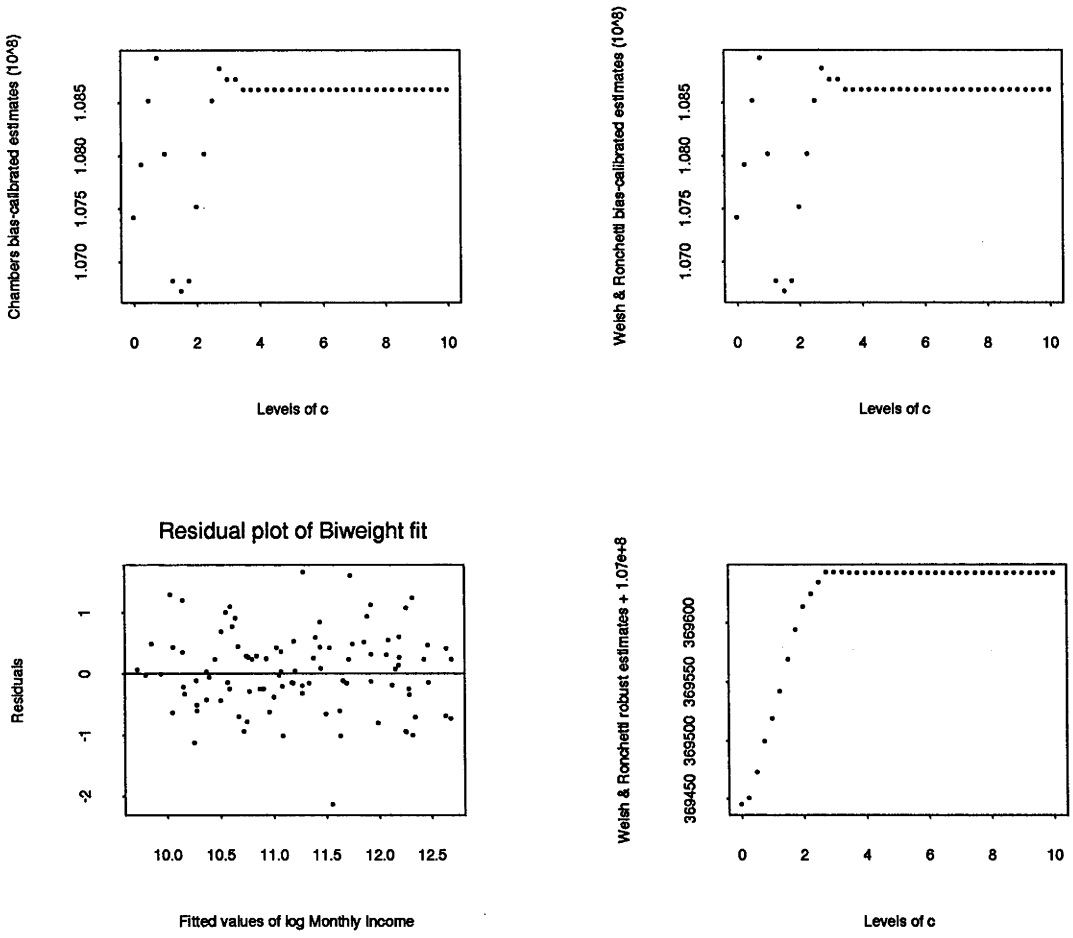


Figure 4.6 : Bias calibrated estimates of Population Total

Sample 3 under Model I

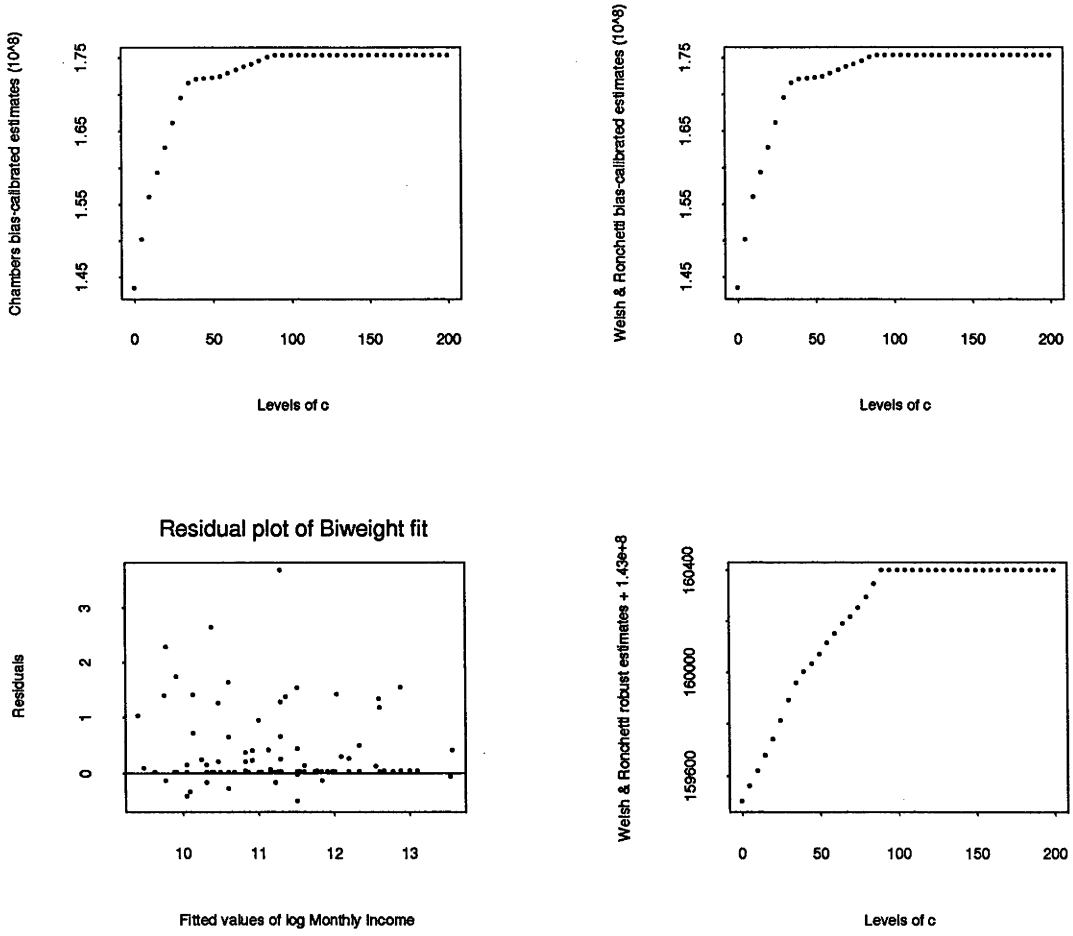
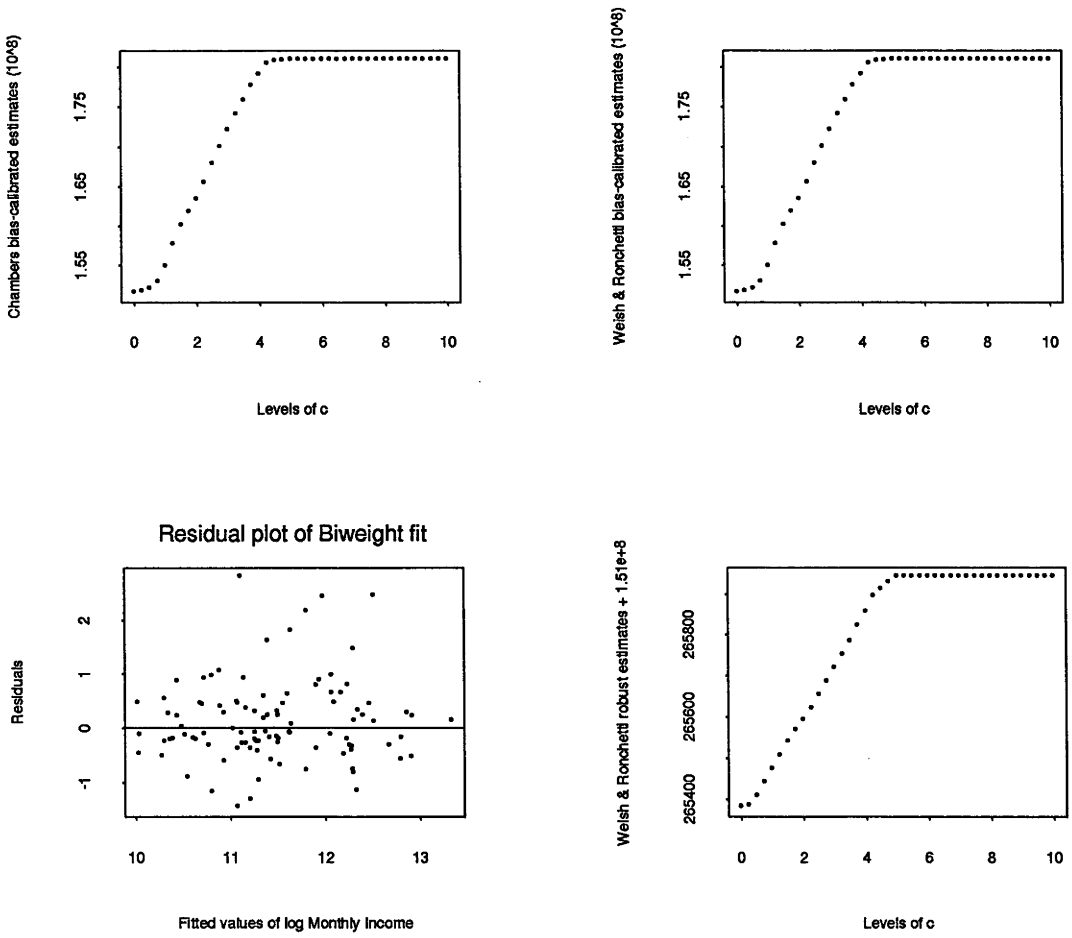


Figure 4.7 : Bias calibrated estimates of Population Total

Sample 3 under Model II



5. Estimating the Population Distribution Function of the Brazilian Data

5.1 *Various Estimators under the Superpopulation Model of the Brazilian Data*

Because of the variance-covariance structure of e under superpopulation models (2.3) and (2.4) is just $\sigma^2 I$, various estimators of distribution function described in chapter can be simplified as the following.

In general, $\hat{F}(t) = \frac{n}{N} F_1(t) + \frac{N-n}{N} \hat{F}_2(t)$

(1) Sample distribution function (3.5)

$$\hat{F}(t) = F_1(t) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \le t) \tag{5.1}$$

(2) Least squares estimator (3.6)

$$\hat{\beta}_{LS} = (X_1' X_1)^{-1} X_1' Y_1$$

$$\hat{F}_{2,LS}(t) = \frac{1}{N-n} \sum_{i=n+1}^{N} I(x_i \hat{\beta}_{LS} \le t) \tag{5.2}$$

(3) Biweight estimator (3.7)

$$\hat{\beta}_R = (X_1' M X_1)^{-1} X_1' M Y_1$$

$$\hat{F}_{2,R}(t) = \frac{1}{N-n} \sum_{i=n+1}^{N} I(x_i \hat{\beta}_R \le t) \tag{5.3}$$

(4) Bias calibrated estimator (3.8)

$$\hat{\beta}_C(c) = \hat{\beta}_R + c \cdot \hat{\sigma} \cdot (\mathrm{X}_1{}' \mathrm{X}_1)^{-1} \mathrm{X}_1{}' \Psi\left\{\frac{\mathrm{Y}_1 - \mathrm{X}_1 \hat{\beta}_R}{c \cdot \hat{\sigma}}\right\}$$

$$\hat{F}_{2,C}(t,c) = \frac{1}{N-n} \sum_{i=n+1}^{N} I(x_i \hat{\beta}_C(c) \le t) \tag{5.4}$$

Note that when $c = 0$, $\hat{F}_{2,C}(t,c) = \hat{F}_{2,R}(t)$ and when $c = \inf$, $\hat{F}_{2,C}(t,c) = \hat{F}_{2,LS}(t)$.

(5) Chambers and Dunstan least squares estimator (3.9)

$$\hat{F}_{2,CD-ls}(t) = \frac{1}{n(N-n)} \sum_{j=n+1}^{N} \sum_{i=1}^{n} I\left\{x_j \hat{\beta}_{LS} + \left(y_i - x_i \hat{\beta}_{LS}\right) \le t\right\} \tag{5.5}$$

(6) Welsh and Ronchetti biweight estimator (3.13)

$$\hat{F}_{2,WR-bi}(t) = \frac{1}{n(N-n)} \sum_{j=n+1}^{N} \sum_{i=1}^{n} I\left\{x_j \hat{\beta}_R + \left(y_i - x_i \hat{\beta}_R\right) \le t\right\} \tag{5.6}$$

(7) Welsh and Ronchetti bias calibrated estimator (3.11)

$$\hat{F}_{2,WR-bc}(t,c) = \frac{1}{n(N-n)} \sum_{j=n+1}^{N} \sum_{i=1}^{n} I\left\{x_j \hat{\beta}_C(c) + \left(y_i - x_i \hat{\beta}_C(c)\right) \le t\right\} \tag{5.7}$$

Obviously, $\hat{F}_{2,CD-ls}(t) = \hat{F}_{2,WR-bc}(t,\infty)$ and $\hat{F}_{2,WR-bi}(t) = \hat{F}_{2,WR-bc}(t,0)$

(8) Welsh and Ronchetti robust estimator (3.15)

$$\hat{F}_{2,WR-r}(t,c) = \frac{1}{n(N-n)} \sum_{j=n+1}^{N} \sum_{i=1}^{n} I\left\{x_j \hat{\beta}_R + c \cdot \hat{\sigma} \cdot \Psi\left[\frac{y_i - x_i \hat{\beta}_R}{c \cdot \hat{\sigma}}\right] \le t\right\} \tag{5.8}$$

When $c = 0$, $\hat{F}_{2,WR-r}(t,c) = \hat{F}_{2,R}(t)$ the biweight estimator and when $c = \infty$, the Welsh and Ronchetti robust estimator reduces to the Welsh and Ronchetti biweight estimator, that is $\hat{F}_{2,WR-r}(t,c) = \hat{F}_{2,WR-bi}(t)$.

5.2 *Performance of the various Distribution Function Estimators on the Brazilian Data*

The various estimators of distribution function described so far are applied to the three selected samples obtained in chapter 4. The performance of each estimator is presented.

Table 5.1 shows the quantile estimates from all estimators based on Sample 1 under model I. For bias calibrated type estimators, the optimal choice of c is found to be the same as the one in chapter 4. Note that abbreviations are used to represent some estimators such that popn=Population, ls=least squares, CDls=Chambers and Dunstan least squares, bi=biweight, WRbi=Welsh and Ronchetti biweight, c1128=bias calibrated with c as 1128, WR1128=Welsh and Ronchetti with c as 1128, r1357=Welsh and Ronchetti robust with c as 1357.

Table 5.1: Quantile estimates of Sample 1 under Model I

quantile	popn	sample	ls	CDls	bi	WRbi	c1128	WR1128	r1357
0.10	20,000	25,000	19,455	18,986	20,000	20,000	19,843	18,990	20,000
0.20	31,560	35,000	32,761	32,436	31,800	31,792	32,864	32,880	31,741
0.30	50,000	50,000	51,192	50,400	50,011	50,600	51,924	51,598	50,600
0.40	64,708	60,000	65,022	65,687	62,014	62,669	64,678	64,661	62,669
0.50	81,270	74,500	84,935	84,935	80,028	81,016	85,291	85,257	81,014
0.60	101,905	88,000	106,751	105,880	100,036	101,400	106,526	106,114	101,399
0.70	150,000	120,000	145,865	145,198	150,000	150,000	145,854	147,176	150,000
0.80	201,997	180,000	209,188	208,520	200,058	202,657	209,649	208,807	202,657
0.90	350,000	301,000	328,519	328,012	300,650	300,601	327,625	327,966	300,601
0.95	500,000	500,000	495,869	492,269	486,108	496,056	500,114	501,381	496,056

Based on this Sample 1, all quantile estimators yield very good estimates for lower quantiles. When higher quantiles are also considered, least squares type estimators are the best. Figure 5.1 shows quantile estimates by the Welsh and Ronchetti bias calibrated estimator and

the Welsh and Ronchetti robust estimator. The Welsh and Ronchetti bias calibrated estimator with c equals to 1128 corresponds to the least squares estimator and the Welsh and Ronchetti robust estimator with c as 0 corresponds to the biweight estimator. It is clear that these two estimated distribution functions do not deviate much from one and other.

Results of Sample 1 under Model II are tabulated below

Table 5.2: Quantile estimates of Sample 1 under Model II

quantile	popn	sample	ls	CDls	bi	WRbi	c21	WR21	r2
0.10	20,000	25,000	15,803	16,030	28,694	28,502	15,803	16,030	28,500
0.20	31,560	35,000	25,987	25,797	40,792	40,694	25,987	25,797	40,693
0.30	50,000	50,000	38,573	38,649	55,716	55,367	38,573	38,649	55,365
0.40	64,708	60,000	54,498	54,846	66,321	66,351	54,498	54,846	66,341
0.50	81,270	74,500	70,640	70,840	79,592	79,357	70,640	70,840	79,352
0.60	101,905	88,000	100,185	100,144	106,003	104,436	100,185	100,144	104,432
0.70	150,000	120,000	165,813	166,457	142,274	142,578	165,813	166,457	142,544
0.80	201,997	180,000	226,123	222,374	197,397	196,864	226,123	222,374	196,862
0.90	350,000	301,000	313,477	312,096	282,258	283,908	313,477	312,096	283,872
0.95	500,000	500,000	388,265	414,710	359,499	372,696	388,265	414,710	372,685

Again Figure 5.2 illustrates that the least squares quantile (the best) estimates do not deviate much from the biweight (the poorest) quantile estimates. Here, c21 and WR21 both correspond to the least squares estimates and r2 corresponds to the biweight quantile estimates. Note that under Model II, all the quantile estimates are also quite satisfactory especially for lower quantiles and major underestimation only occur from 0.85 quantile.

Table 5.3: Quantile estimates of Sample 2 under Model I

quantile	popn	sample	ls	CDls	bi	WRbi	c0	WR0	r10
0.10	20,000	18,100	20,666	20,427	19,693	18,744	19,693	18,744	18,789
0.20	31,560	30,000	32,262	32,100	30,044	31,029	30,044	31,029	31,098
0.30	50,000	40,000	46,138	46,866	49,967	49,026	49,967	49,026	49,098
0.40	64,708	60,000	61,090	60,608	60,100	60,182	60,100	60,182	60,183
0.50	81,270	70,635	74,698	74,830	80,088	80,666	80,088	80,666	80,671
0.60	101,905	100,000	91,614	91,053	100,103	100,309	100,103	100,309	100,313

0.70	150,000	130,000	114,971	114,944	136,470	136,993	136,470	136,993	136,995
0.80	201,997	200,000	155,887	155,775	200,132	200,310	200,132	200,310	200,293
0.90	350,000	305,600	225,723	225,393	300,371	304,367	300,371	304,367	304,373
0.95	500,000	405,000	305,965	308,175	486,311	491,340	486,311	491,340	489,514

Due to the negative outliers of Sample 2, least squares type estimators are seriously affected (Figure 5.3). In this case the biweight type estimators have the best performance. Here, the Welsh and Ronchetti robust estimator (c=10) is very similar to the biweight type estimators.

Table 5.4: Quantile estimates of Sample 2 under Model II

quantile	popn	sample	ls	CDls	bi	WRbi	c3	WR3	r3
0.10	20,000	18,100	26,057	25,888	24,407	24,570	26,074	25,913	24,570
0.20	31,560	30,000	38,985	38,999	35,343	35,654	38,653	38,535	35,654
0.30	50,000	40,000	50,556	50,733	50,681	50,550	50,810	50,507	50,550
0.40	64,708	60,000	60,629	60,704	60,246	60,469	60,398	60,813	60,469
0.50	81,270	70,635	72,761	73,658	73,661	74,007	74,375	74,126	74,007
0.60	101,905	100,000	98,352	96,422	100,000	100,000	98,698	97,641	100,000
0.70	150,000	130,000	127,746	128,719	132,804	132,519	129,125	129,497	132,519
0.80	201,997	200,000	177,772	180,026	177,729	177,678	177,227	179,325	177,678
0.90	350,000	305,600	256,923	260,983	250,590	250,275	256,973	258,986	250,275
0.95	500,000	405,000	325,575	331,260	311,355	310,973	324,764	330,986	310,973

For Sample 2 under Model II, estimates yielded by the least squares type estimators are basically the same as what the biweight type estimators yield. Yet in this case severe underestimation occurs for quantiles 0.7 and above. Note that in this situation bias calibrated type estimators have very little effect.

Table 5.5: Quantile estimates of Sample 3 under Model I

quantile	popn	sample	ls	CDls	bi	WRbi	c3	WR3	r3
0.10	20,000	26,000	29,265	29,698	20,289	20,153	20,909	20,167	20,153
0.20	31,560	40,000	44,691	48,223	33,279	33,159	33,854	33,950	33,159
0.30	50,000	60,000	702,73	70,323	50,475	50,565	52,494	52,973	50,565
0.40	64,708	73,270	849,66	89,297	64,439	64,471	66,636	66,838	64,471

0.50	81,270	100,000	110,000	109,999	85,280	83,796	87,364	89,051	83,796
0.60	101,905	120,000	136,423	137,087	100,612	101,942	105,539	108,612	101,942
0.70	150,000	150,635	191,228	181,275	150,000	150,000	153,148	151,816	150,000
0.80	201,997	300,000	254,497	264,555	199,833	197,051	207,240	216,418	197,051
0.90	350,000	450,000	401,003	405,172	347,986	338,451	359,164	342,391	338,451
0.95	500,000	700,000	592,399	590,413	495,949	467,658	510,950	515,793	467,658

For Sample 3 which has a few of positive outliers under Model I, bias calibrated type estimators become important as least squares type estimators tend to overestimate the true distribution function seriously and the biweight type estimators underestimate it mildly. Figure 5.5 and Figure 5.6 display the effects of bias calibration of the Welsh and Ronchetti bias calibrated estimator and the predicted values augmented from the Chambers's bias calibrated estimator. They do not seem to be different. Hence the adjustment from Chambers' estimator to Welsh and Ronchetti's estimator is not significant for this sample. Figure 5.7 shows that the bias calibration of the Welsh and Ronchetti robust estimator basically has no effect at all for this sample, yet its performance is very close to the best estimator (Welsh and Ronchetti bias calibrated estimator with c=3, see Figure 5.8). Therefore in this situation, if the true distribution is unknown which is true in reality, the Welsh and Ronchetti robust estimator should be chosen to estimate the distribution function.

Table 5.6: Quantile estimates of Sample 3 under Model II

quantile	popn	sample	ls	CDls	bi	WRbi	c3	WR3	r3
0.10	20,000	26,000	27,782	27,983	29,230	28,507	29,230	28,507	28,484
0.20	31,560	40,000	44,540	45,058	42,793	42,715	42,793	42,715	42,815
0.30	50,000	60,000	60,697	60,914	59,475	59,576	59,475	59,576	59,637
0.40	64,708	73,270	78,115	78,453	73,220	73,001	73,220	73,001	72,809
0.50	81,270	100,000	100,000	100,000	94,336	94,359	94,336	94,359	94,241
0.60	101,905	120,000	131,244	131,056	120,222	120,198	120,222	120,198	120,299
0.70	150,000	150,635	187,032	185,662	165,505	165,679	165,505	165,679	165,149
0.80	201,997	300,000	296,741	296,191	251,235	251,120	251,235	251,120	251,019
0.90	350,000	450,000	485,414	477,125	378,417	373,475	378,417	373,475	374,601
0.95	500,000	700,000	640,233	654,018	482,565	485,283	482,565	485,283	483,518

For Sample 3 under Model II, all estimators tend to overestimate the true distribution. Thus, biweight type estimators are preferred. Figure 5.9 shows the difference between least squares estimates and biweight estimates. Even though the bias calibration of the Welsh and Ronchetti robust estimator again has very little effect (Figure 5.10), it is equivalent to the biweight type estimator hence the best estimator.

5.3 *Discussion*

Results of the various distribution function estimators applied to the three samples are consistent with the results of their corresponding total estimators. For sample without outliers, least squares type estimator is the best option (Figure 4.2 and Figure 5.1) but in some circumstances like the situation under the Model II, all model based estimators lead to misleading estimates especially for higher quantiles (Figure 4.5 and Figure 5.4). A suggestion to prevent the adoption of these misleading model based estimates is to compare the quantile estimates given by the sample distribution function and the quantile estimates yielded by the model based estimators. If the two are varied distantly and consistently (especially for higher quantiles) then care should be taken about whether one should rely on the model based estimates. This is illustrated by Table 5.1 and Table 5.4. That is, if both the biweight and least squares estimators yield estimates very different from the sample distribution function estimates for higher quantiles. More research may be needed for dealing with this situation.

For sample with either positive or negative outliers, bias calibrated type estimators are able to give estimates surpassing other type of estimators if the choice of c is appropriate (eg. Figure 5.8). Welsh and Ronchetti (1994) has provided an approximate predictive Bayesian argument for the use of an additive calibration term. This in fact is a pioneer work on the link of the choice of c with the use of prior information available in survey sampling but in practice how the prior information should be and can be used to choose c is still unknown. More research is needed to solve this problem.

Welsh and Ronchetti (1994) has also suggested a new method to improve the quantile estimates which is to use the Welsh and Ronchetti robust estimator with c varying over the support of the distribution in such a way that c increases as we move into the right tail. We

found that in the case of Brazilian data, range of different c of the Welsh and Ronchetti robust estimator is extremely narrow and this is probably due to the variance -covariance structure of e of models (2.3) and (2.4) where $\text{var}(e)=\sigma^2 I$ such that the adjustment term of equation (5.8)

$$\hat{F}_{2,WR-r}(t,c)=\frac{1}{n(N-n)}\sum_{j=n+1}^{N}\sum_{i=1}^{n} I\left\{x_j\hat{\beta}_R + c\cdot\hat{\sigma}\ \cdot\Psi\left[\frac{y_i - x_i\hat{\beta}_R}{c\cdot\hat{\sigma}}\right]\le t\right\}$$

is insignificant. Therefore the Welsh and Ronchetti robust estimator here behaves very much the same as the biweight estimator. However, the non-heterocedasticity of the Brazilian data is rare in sample survey, the usefulness of the Welsh and Ronchetti robust estimator in outlier robust estimation still has much to be explored.

Nonetheless, the idea of using different values of the calibration constant c in different parts of the distribution is supported by this analysis (Figures 5.5, 5.6 and 5.8). With data like the Brazilian data which is not heterocedastic in nature, to estimate the quantiles, the Welsh and Roncheti bias calibrated estimator (or the bais calibrated estimator) is suggested with c varying over the support of the distribution in such a way that c increases as we move into the right tail.

5.4 *A Strategy of Estimating the Population Distribution Function*

I. When the random sample does not contain outliers, least squares estimators should be the first preference when model misspecification is not a problem.

II. When the random sample contains either negative outliers or positive outliers, the Welsh and Ronchetti robust estimator is suggested, especially when there is no heterocedasticity in the data such that the adjustment term of the bias calibrated estimators is insignificant. However, if the data shows heterocedasticity and therefore the adjustment term of the bias calibrated estimators is significant, then the Welsh and Ronchetti robust estimator with c varying over the support of the distribution would be a good option.

Figure 5.1 Plots of the quantile functions of Sample 1 under Model I from the Welsh & Ronchetti's estimators of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.

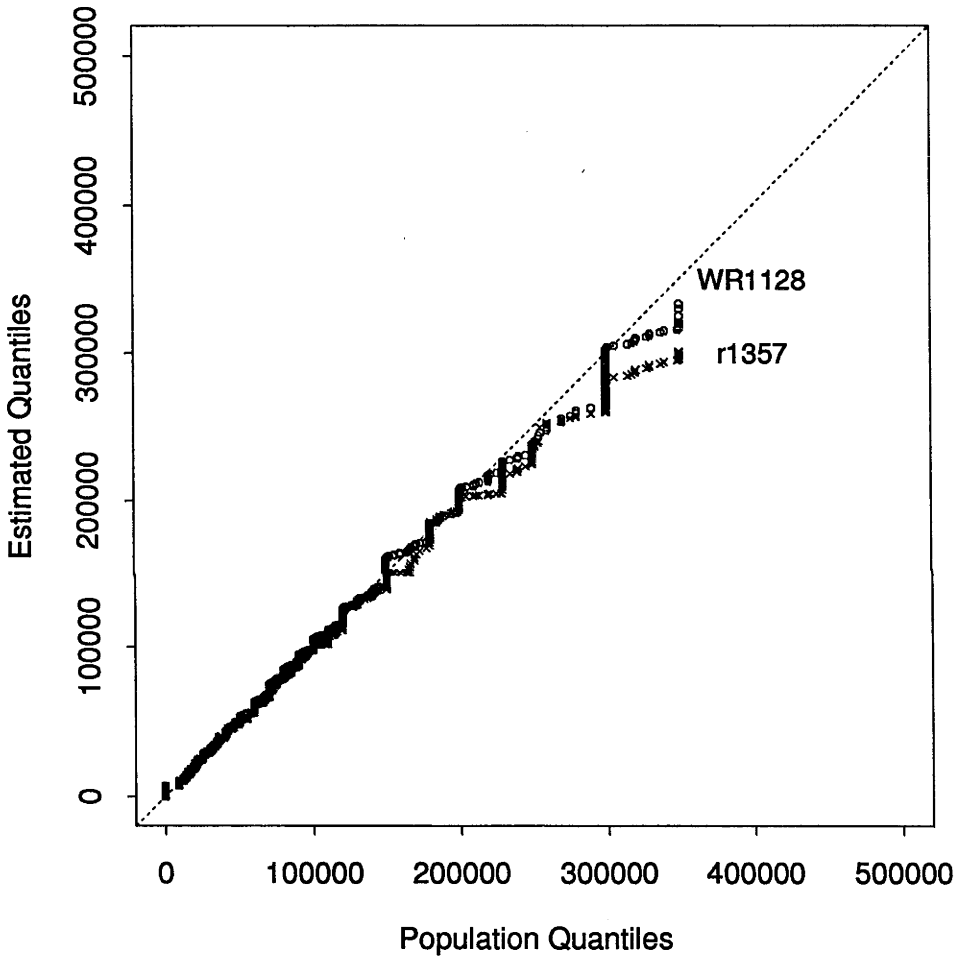


Figure 5.2 Plots of the quantile functions of Sample 1 under Model II from the Welsh & Ronchetti's estimators of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.

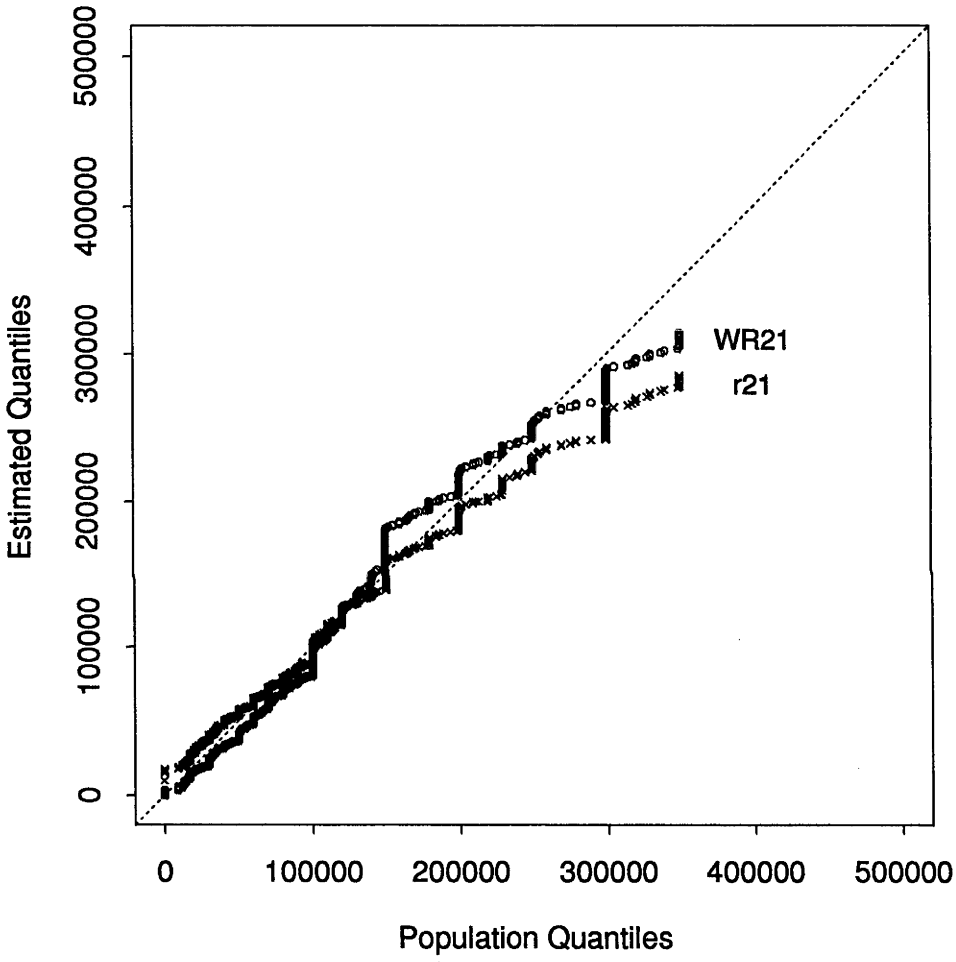


Figure 5.3 Plots of the quantile functions of Sample 2 under Model I from the W & R's estimators and C & D's estimator of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.

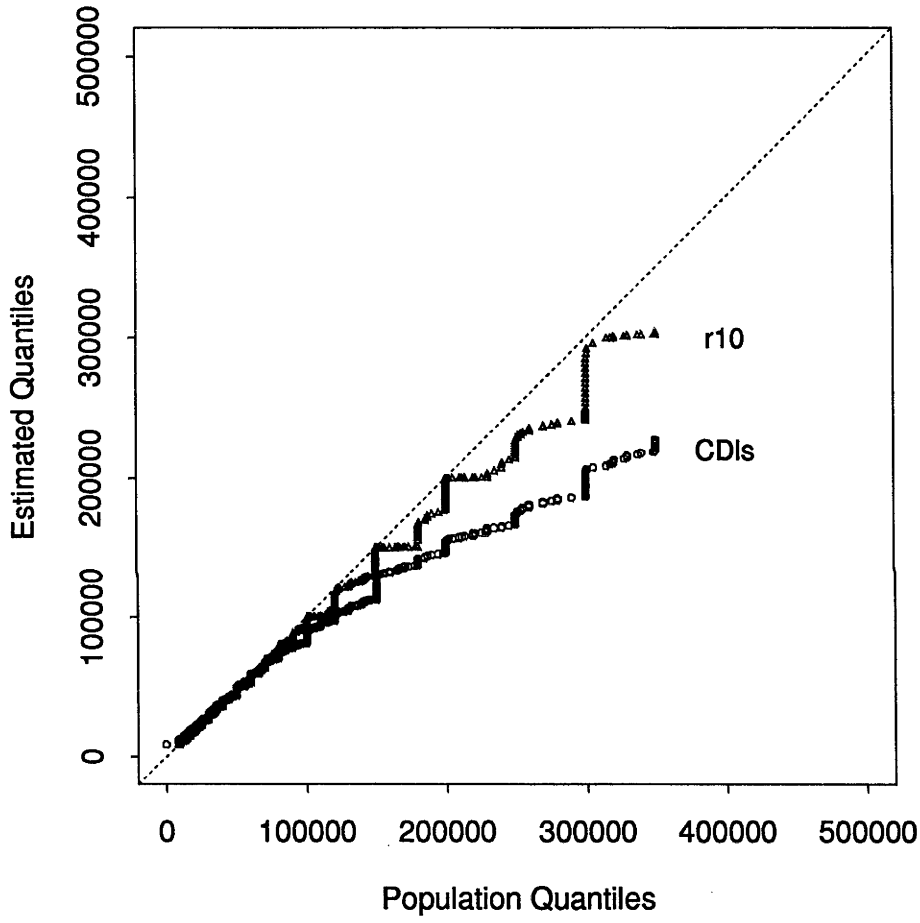


Figure 5.4 Plots of the quantile functions of Sample 2 under Model II from the Welsh & Ronchetti's estimators of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.

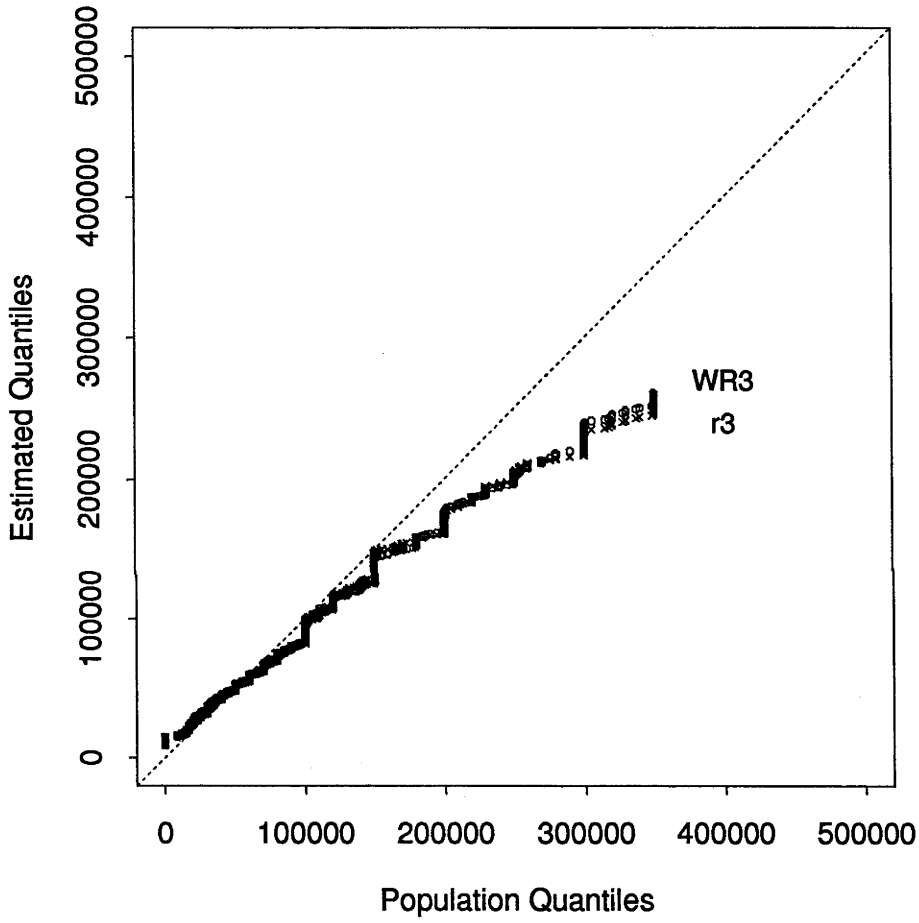


Figure 5.5 Plots of the quantile functions of Sample 3 under Model I augmented by the predicted values from Chambers' estimator of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.

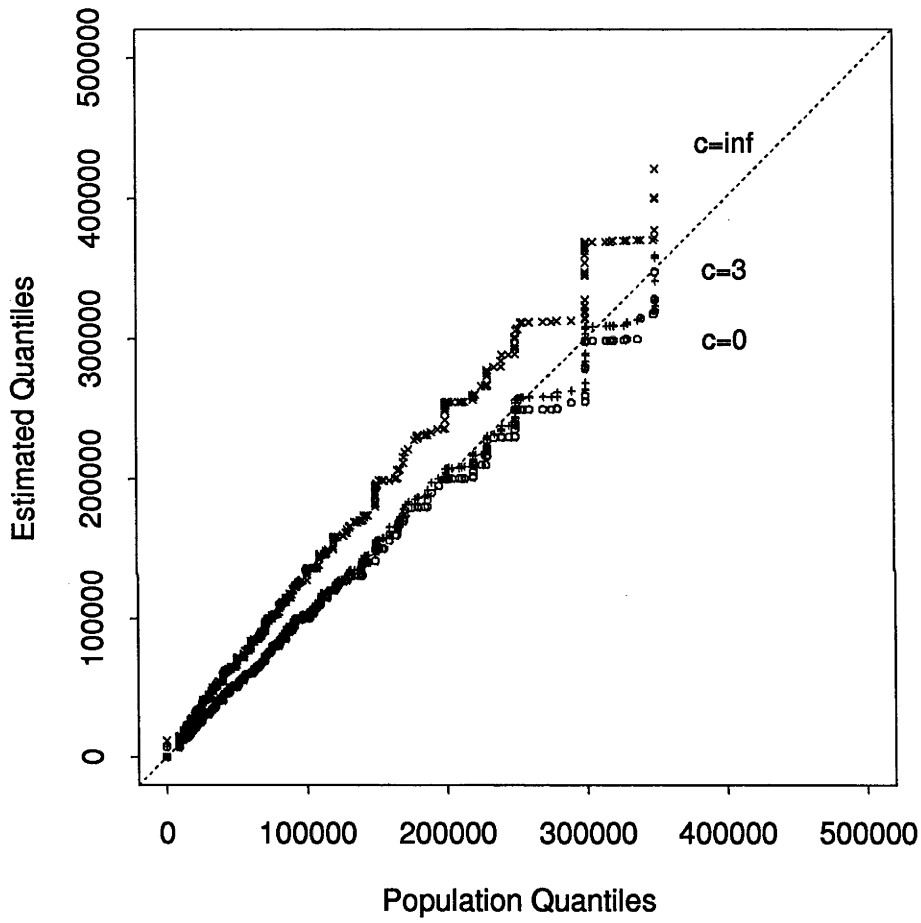


Figure 5.6 Plots of the quantile functions of Sample 3 under Model I from WRbc's estimator of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.

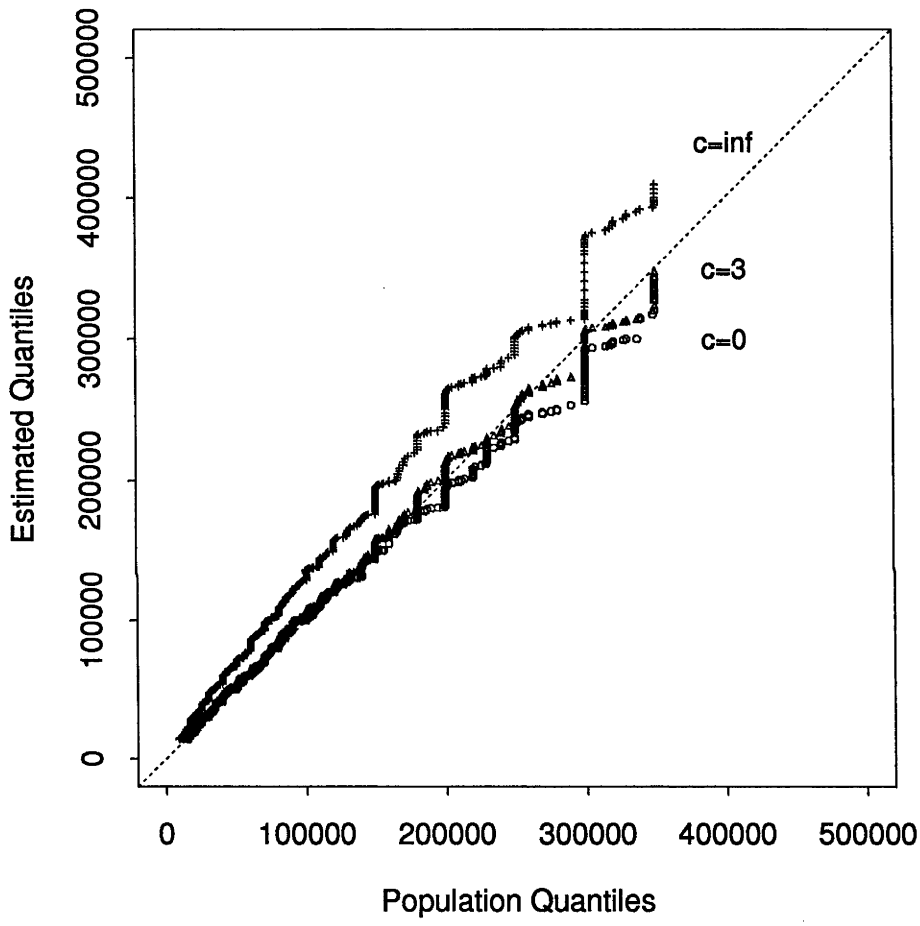


Figure 5.7 Plots of the quantile function of Sample 3 under Model I from WRr's estimator of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.

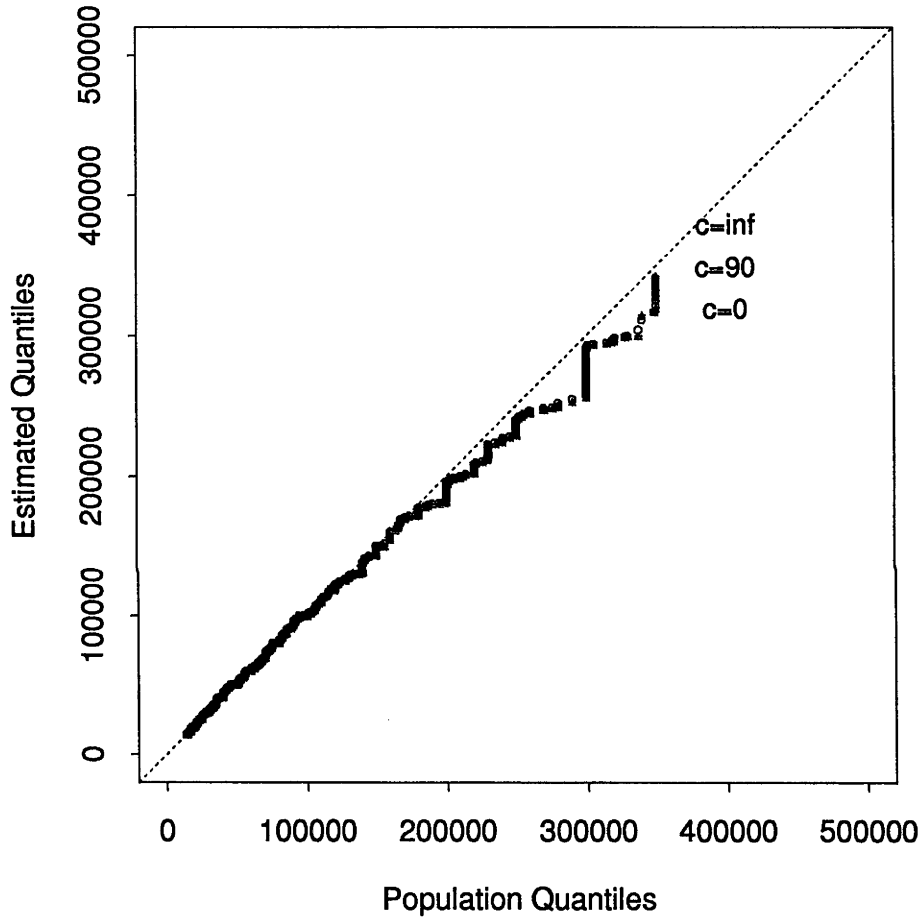


Figure 5.8 Plots of the quantile functions of Sample 3 under Model I from the W & R's estimators and C & D's estimator of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.

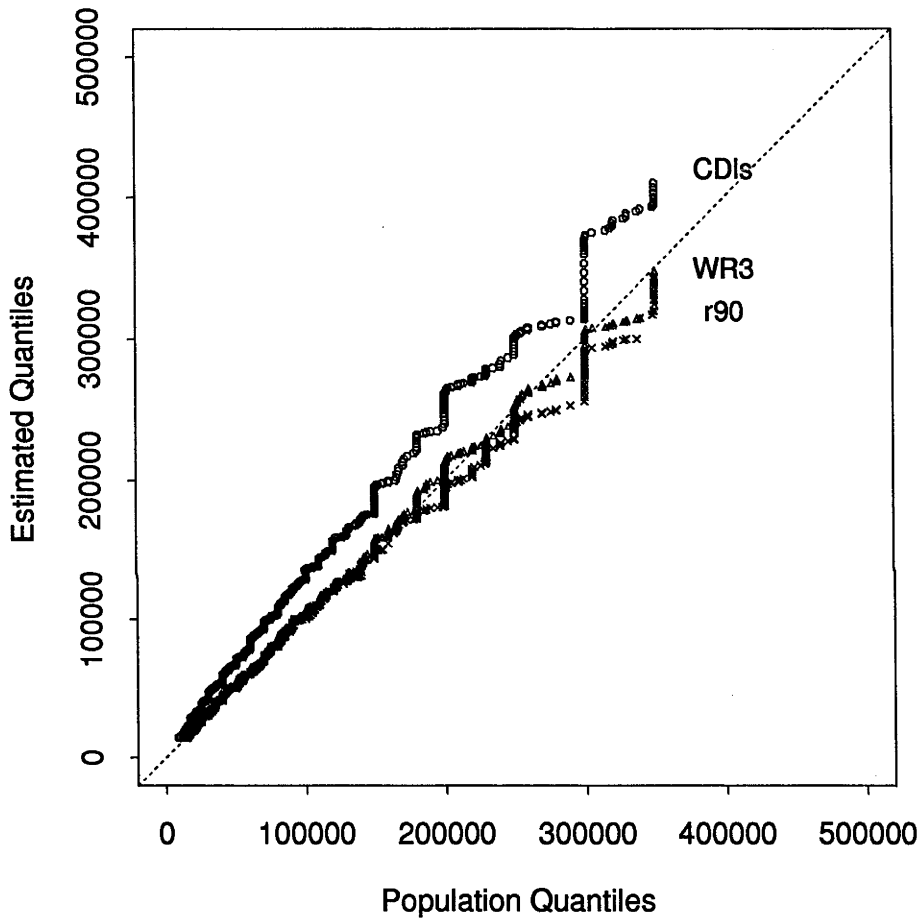


Figure 5.9 Plots of the quantile functions of Sample 3 under Model II by Welsh and Ronchetti bias calibrated estimator of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.

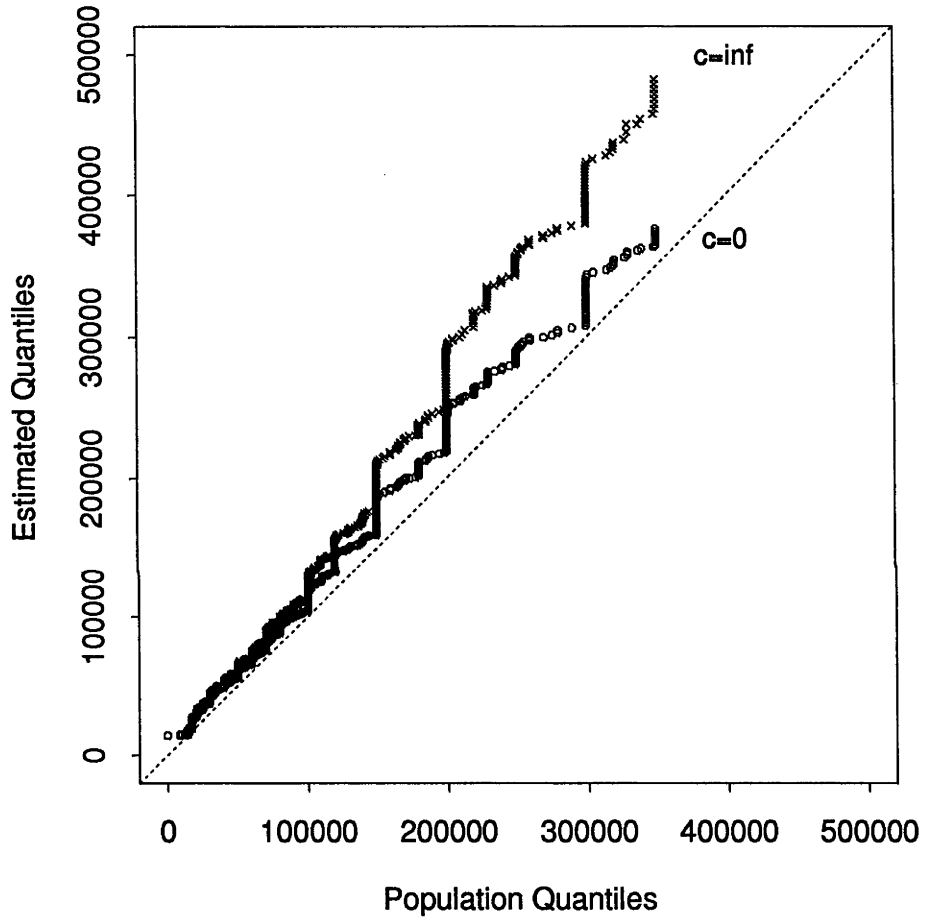
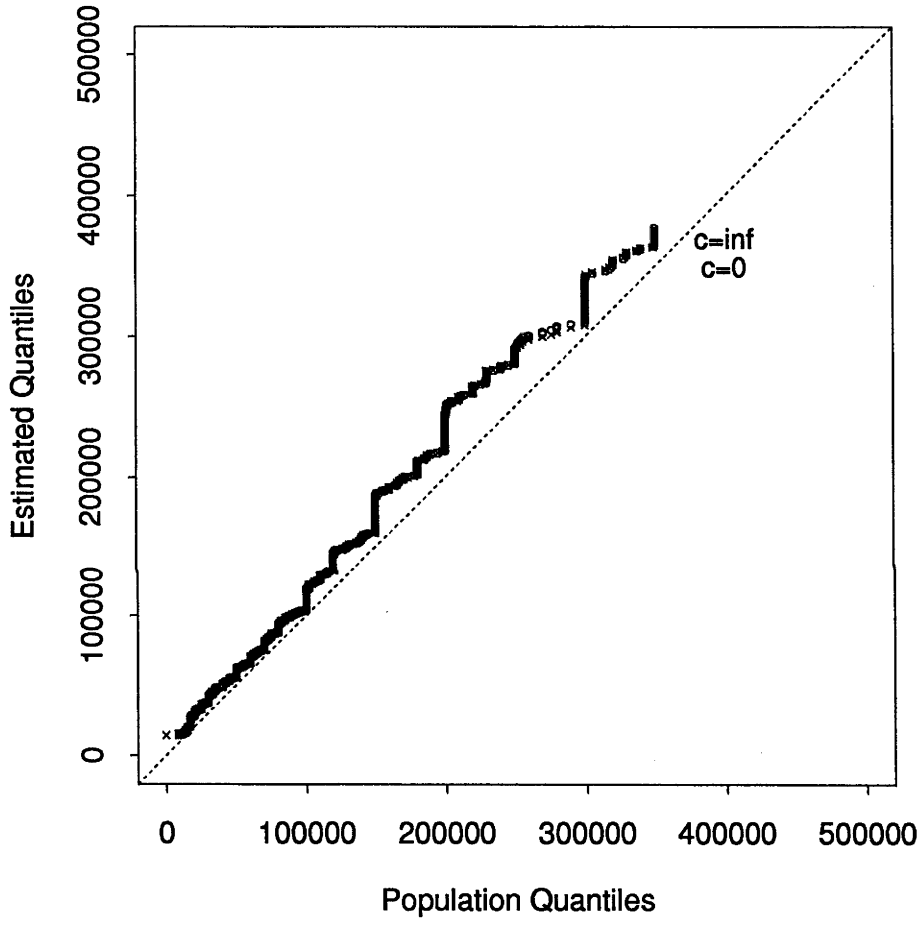


Figure 5.10 Plots of the quantile functions of Sample 3 under Model II by Welsh and Ronchetti robust estimator of the population quantile function. The quantile functions are truncated at the 0.90 quantile to show detail over the range of moderate quantiles.



6. Conclusion

The Brazilian data set used in this thesis is undoubtedly a rare case in survey sampling since it seems that the log transformations of the Monthly Total Income has taken care of problems like heterocedasticity, non-normality etc. As outliers problem is also not serious, the extent of research of the role of bias calibration in finite population estimation in this thesis is limited. Nevertheless, the work of this thesis has revealed to certant extent the nature of the various bias calibrated estimators. Some final remarks are

(1) Bias calibrated estimators are most useful when the random sample contains positive outliers. The usual residual plot of diagnostics check is an important tool to discern whether the outliers are positve, negative or ther are no outliers at all.

(2) In practice, the choice of the optimal c of the bias calibrated estimator is difficult to determine. Thus, in estimating population distribution function, the use of the Welsh and Ronchetti robust estimator with different values of c in different parts of the distribution is very attractive. The robust property of this estimator is also very desirable.

(3) An extention of the use of the Welsh and Ronchetti robust distribution function estimator (with different values of c in different parts of the distribution) to estimate the population total is natural. Unfortunately, the Brazilian data here has limited us to explore this extention. But as long as the problem of the optimal c is unsolved, this way to estimate the true total seems to be the supreme one.

REFERENCES

Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band spectroscopic data. *Technometrics* **16** 147-185.

Brewer, K. R. W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Austral. J. Statist.* **5** 93-105.

Chambers, R. L. (1986). Outlier robust finite population estimation. *J. Amer. Statist. Assoc.* **81** 1063-1069.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73** 597-604.

Chambers, R. L. and Kokic, P. (1993) Outlier robust sample survey inference. In *Bulletin of the International Statistical Institute: Proceedings of the 49th Session, Book 2.*

Chambers, R. L., Dorfman, A. H. and Wehrley, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *J. Amer. Statist. Assoc.* **88** 268-277.

Glasser, G. J. (1962). On the complete coverage of large units in a statistical study. *Intl. Statist. Instit. Reviews* **30** 28-32.

Hidiroglou, M. H. and Srinath, K. P. (1981). Some estimators of the population total from simple random samples contaning large units. *J. Amer. Statist. Assoc.* **76** 690-695.

Kish, L. (1965) *Survey Sampling* Wiley, New York.

Rao, C. R. (1971). Some aspects of statistical inference in problems of sampling from finite populations. In *Foundations of Statistical Inference* (eds V. P. Godambe and D. A. Sprott), Holt, Rinehart and Winston, Toronto.

Searls, D. T. (1966). An estimator which reduces large true observations. *J. Amer. Statist. Assoc.* **61** 1200-1204.

Welsh, A. H. and Ronchetti, E. (1994). Bias-calibrated estimation from sample surveys containing outliers. Unpublished manuscript.